Altruistic Motivation Beyond Altruistic Desires

Jorge Piaia Mendonca Junior (B.A., M.A.)

This thesis is presented for the degree of Doctor of Philosophy of

The University of Western Australia

School of Humanities

Discipline of Philosophy

2023

Thesis Declaration

I, Jorge Piaia Mendonca Junior, certify that:

This thesis has been substantially accomplished during enrolment in this degree.

This thesis does not contain material which has been submitted for the award of any other degree or diploma in my name, in any university or other tertiary institution.

In the future, no part of this thesis will be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of The University of Western Australia and where applicable, any partner institution responsible for the joint-award of this degree.

This thesis does not contain any material previously published or written by another person, except where due reference has been made in the text.

This thesis does not violate or infringe any copyright, trademark, patent, or other rights whatsoever of any person.

This thesis does not contain work that I have published, nor work under review for publication.

Signature: 

Date: 01/05/2023

Abstract

The term “altruism” is used in many ways. In this thesis, I discuss altruism as a *motivation*, which is an influential notion in philosophy and the social sciences. Questions about the nature and the possibility of altruistic motivation have inspired much debate, both in academia and in everyday conversations. How can we know when we are truly altruistic and when we are merely helping others as a means to some egoistic goal? Are humans even capable of genuine altruistic motivation or are we always egoistic deep down? Before answering questions like these, however, we need to ask a more fundamental question: how should we *define* altruistic motivation? The standard account of altruistic motivation in the literature defines it as an ultimate desire to increase the welfare of others. The central claim I will argue for in this thesis is that this standard account is intrinsically flawed and should be abandoned.

The problem with the standard account of altruistic motivation, in short, is its reliance on the notion of “ultimate desire”, which is a limited and unfruitful way of conceptualizing altruistic motivation. The criticism I propose is based on an interdisciplinary analysis of many aspects of the standard account, discussing its philosophical basis, its use in scientific research, its historical development, and its relation to common sense and morality.

This thesis is divided into four parts, each composed of two chapters. Part I presents the many technical accounts of altruism present in the scientific literature, distinguishing the standard account of altruistic motivation from alternative accounts of altruism. I discuss the hypotheses of psychological altruism, which states that *some* of our ultimate desires aim to increase the welfare of others, and psychological egoism, which states that *all* of our ultimate desires are egoistic. In this first part, I discuss some of the central philosophical concepts

underlying the standard account of altruistic motivation, showing their implications for the debate about psychological altruism.

Part II discusses the main arguments for the existence of altruistic motivation. I argue that the main arguments in the scientific literature fail in making a defense of psychological altruism and that the definition of altruistic motivation makes the solution to the debate on psychological altruism virtually impossible. I conclude that the standard account makes altruistic motivation an unverifiable and unfruitful notion for scientific research.

Part III discusses the history of altruism and egoism, focusing particularly on early modern philosophy. I argue that the modern accounts of altruism and egoism diverge from the contemporary standard account of altruistic motivation. Reading the debates in modern philosophy through the lens of psychological altruism and psychological egoism oversimplifies and impoverishes the modern debate. I argue that the standard account also fails in representing the original use of "altruism", proposed by Auguste Comte. In this third part, I show how the use of the standard account of altruistic motivation cannot be based on its historical foundations.

Finally, Part IV addresses the normative character of altruism and discusses some alternative ways of conceptualizing altruism. I argue that the standard account of altruistic motivation does not reflect the ordinary use of the term "altruism" and fails in accounting for its normative character. As an alternative, I propose an alternative account of altruism, *virtue altruism*. I claim that this account better reflects the ordinary intuitions about altruism, accommodates the descriptive and normative dimensions of altruism, is at home with most of the historical accounts of altruism, and may promote the active promotion of altruistic actions.

Table of Contents

Acknowledgments

Firstly, I would like to express my deepest gratitude to my advisor, Prof. Robert A. Wilson, for all the patience, myriad comments on my drafts, and overall support throughout these years. To my co-advisor, Dr. Remco Heesen, my sincere thanks for all the profoundly helpful comments on my drafts, always delivered in the kindest way possible. Thanks also to Dr. Richard Heersmink, who was my former co-advisor in the early stages of this thesis. I am also indebted to my former advisors, Prof. Nythamar de Oliveira and Prof. John Sarnecki, who taught me so much and helped me to progress in my career.

This research was supported by a Scholarship for International Research Fees (SIRF) and an International Living Allowance Scholarship (Ad Hoc Postgraduate Scholarship). In the first 17 months, this research was funded by a La Trobe University Postgraduate Research Scholarship (LTUPRS) and a La Trobe University Full-Fee Research Scholarship (LTUFFRS). I am truly grateful for all this support.

Thanks to my friends and colleagues, Alberto Guerrero and Lucia Neco, for all the help in the process of developing the ideas present in this thesis and for being a home-like presence in a place far from home. I also express my sincere appreciation for my colleagues and coworkers who treated me so well, making my life in Australia much easier: Adam Andreotta, Alan Tapper, Anne Schwenkenbecher, Chris Letheby, Genevieve Hawks, Ian Maddox, Jacqueline Boaks, Kaz Bland, Lachlan Umbers, Michael Rubin, Miri Albahari, Phil Pickering, Richard Hamilton, and Simon Kidd. I learned a lot from all of you. Thanks also to my friends in Brazil and the USA, who have encouraged me since the beginning of this project: Ayune Veríssimo, Claiton Costa, Dani Quinsani, Jordan Cook, Marcelo Bujak, Mari Gil, and Will Delzeith.

Finally, my special thanks to my brother, Gabriel, for being my main inspiration to be a better person, and to Li Xi, for all the love and patience throughout these years. I could not have undertaken this journey without you two.

Authorship Declaration: Sole Author Publications

This thesis does not contain published material nor material prepared for publication.

Signature: 

Date: 01/05/2023

Chapter 1

Introduction

Humans are not always nice to each other. But sometimes we see people giving money to charity, volunteering, and even donating organs to complete strangers. We often use the word "altruism" to refer to these cases. But what does "altruism" mean, exactly? There is no simple answer to this question.

The French word "*altruisme*" first appeared in the writings of the founder of Positivism, Auguste Comte[1] (1851/1875, Vol. 1). The neologism "*altruisme*" was derived from the Italian "*altrui*", "meaning 'of or to others, what is another's, somebody else'", which "in turn, was based on the Latin *alteri huic*, meaning 'to this other'" (Dixon, 2008, p. 19). In Comte's work, altruism was conceived as the opposite of egoism, referring to "selfless or other-regarding instincts, emotions, or motives" (Dixon, 2008, p. 4). Since Comte, "altruism" has been assimilated by both scientific and colloquial vocabularies, evolving into several different forms. There is no single answer to the question of what altruism means: the range of phenomena described by "altruism" goes from the higher-order humanitarian motivation of someone donating a kidney to a stranger to the aggregating behavior of slime molds.

A starting point to making sense of the "many faces" of altruism is to distinguish two kinds of altruism. The first takes altruism to be a kind of *behavior*. In this behavioral account of altruism, we disregard the motives that agents may have to help others and apply the label

[1] Although it has been common in the literature to say that Comte coined the term, some sources claim that altruism was used before by one of Comte's teachers, François Andrieux. One of these sources, who was a follower of Comte, claims that Comte himself affirmed that "altruism" was Andrieux's creation (Dixon, 2008, p. 47).

"altruistic" to certain behaviors. By contrast, a second kind depicts altruism as a *motivation* underlying helping behaviors. In this motivational account, altruism refers to the motivational states that make us help others. This thesis focuses on the latter account, that is, altruism as a motivation.

Claims that a particular act was motivated by genuine altruistic motivation commonly face suspicion. A common view throughout the history of philosophy and the social sciences is that humans act mainly out of egoistic motivation. For any helping behavior, the explanation that such behavior is a means to an egoistic end is traditionally regarded as the simpler hypothesis. The egoistic end can be an external reward, such as social praise, or something internal, such as the good feeling one has after helping. This widespread acceptance of the dominance of egoistic motivation cultivated a strong skepticism toward the idea that humans are capable of genuinely caring for each other without expecting anything in return. Those who are inclined to believe in altruistic motivation are often accused of doing so because, deep down, "they want the world to be a friendly and hospitable place" (Sober & Wilson, 1998, p. 8). Rejecting the existence of altruistic motivation has been a safe strategy to dissociate oneself from what others may regard as a naïve worldview. However, many authors have resisted the egoistic explanation of human motivation. Some argue that, although humans are often egoistic, there is still room for altruistic motivation in our motivational repertoire.

In the contemporary literature, the debate opposing altruistic and egoistic motivations is framed in terms of two hypotheses. The hypothesis of *psychological altruism* states that *some of our ultimate desires aim to increase the welfare (or wellbeing) of others* (Sober & Wilson, 1998; Stich et al., 2010). This is a counterpoint to another hypothesis, known as *psychological egoism*, which states that *all of our ultimate desires aim to increase our own welfare* (Feinberg, 2013;

Sober, 2013). The issue under dispute, thus, is whether altruistic ultimate desires exist or whether humans are always, deep down, following their egoistic ultimate desires. The question about whether psychological altruism is true is what Batson (1991) calls the "altruism question" (p. vii).

Empirical research plays an important role in the contemporary literature discussing the altruism question. Rather than following an aprioristic approach, the philosophical arguments presented for psychological altruism are informed by (or based on) scientific research. This scientifically oriented approach has produced a rich and markedly interdisciplinary literature, which will be discussed in this thesis (e.g., Batson, 1991, 2011; Schulz, 2016, 2018; Sober, 2013; Sober & Wilson, 1998; Stich, 2007; Stich et al., 2010). In this contemporary debate, altruistic motivation is defined as an ultimate desire to increase the welfare (or wellbeing) of others. This account of altruism, which I will call the "standard account of altruistic motivation", is the main object of study in this thesis.

Different from most authors engaged in this debate, my goal is *not* to argue for or against the existence of altruistic motivation. I am concerned with a more fundamental question. Philosophers and scientists have accepted the reduction of altruistic motivation to ultimate desires to increase the welfare of others. But do we have good reasons to adopt this particular account of altruistic motivation? In this thesis, I will argue for a negative answer to this question. This interdisciplinary project explores many dimensions of this account of altruism, assessing the different ways in which it is used, and showing how this way of thinking about altruism is problematic and should be abandoned. Each of the four parts of this thesis offers a different reason to support this criticism.

Before giving a summary of the four parts of this thesis, I will present some reasons why our definition of altruistic motivation and the debate on psychological altruism are *relevant*. Firstly, the debate on psychological altruism has a purely theoretical relevance: the altruism question says something about human psychology, and independently of which particular use it has, it tells us something about the human mind (Sober & Wilson, p. 273). But I want to highlight consequences that go beyond this purely theoretical interest.

Psychological altruism and psychological egoism are merely *descriptive*, not *normative* hypotheses. However, the answer to the altruism question has implications for how we see ourselves and for how we see others. There are some subtle ways in which this can affect our behaviors. The modern philosopher Francis Hutcheson (1694–1746) already urged that philosophers should not only discuss whether different accounts of human nature are *true or false*, but also take into account the *implications* of holding these accounts. The actions people take in the world are influenced by what they believe about human nature, and Hutcheson (1728/2002) argued that accounts of human nature that depict humans as egoistic creatures have pernicious effects, discouraging one's cultivation of one's "*kind generous Affections*" (p. 3).

Following Hutcheson, I claim that believing in psychological egoism has the potential to work as a self-fulfilling prophecy, where the belief that one's motivations are egoistic encourages one's egoistic motivations. A believer of psychological egoism might have good reasons to perceive activities such as volunteering as inherently hypocritical or based on self-delusion. If they do so, then it seems likely that they will be discouraged from pursuing altruistic behaviors. If they believe that all their motivations are ultimately egoistic, whenever they feel a desire to help, even if such a desire seems to be altruistic to them, they will know that it is egoistic deep down (see Broadie & Smith, 2022). A grey filter is added to their vision, where all

acts of kindness, all small gestures of caring, appear as egoistic actions based on dissimulation or pretense. To those who are averse to the idea of seeing themselves as dissimulate or hypocritical, acting altruistically would be discouraged by the belief in psychological egoism (see Broadie & Smith, 2022).

Another important implication of the debate on psychological altruism concerns its moral implications. This debate is not only part of a fundamental discussion in moral philosophy, but, as Clavien (2012) claims, many people consider that "an action is morally good only if it is caused by non self-interested motives" (p. 277). For these people, psychological egoism could simply undermine the rational justification for avoiding egoistic actions, since it entails that all of our actions are necessarily based on egoistic desires anyway. It seems that Lichtenberg (2008) is right in pointing out that psychological egoism "provides a convenient excuse for selfish behavior" (p. 4). The reasoning is that, if psychological egoism is true, then even the *apparently* altruistic desires that we may have are ultimately egoistic. "If 'everybody is like that' [egoistic]… we need not feel guilty about our own self-interested behavior or try to change it" (Lichtenberg, 2008, p. 4). One might be better off simply embracing the unavoidable egoism of the human condition.

But whether or not believing in psychological egoism, in fact, discourages people's helping behaviors is, of course, an empirical question. Interestingly, there is empirical support for this claim. Frank et al. (1993) show that students of economics tend to behave more egoistically than people from different areas. In prisoner's dilemma games, for example, economics majors defect more than noneconomic majors (Frank et al., 1993, p. 165). The authors argue that these results indicate an effect of the students' training in economics, since this area commonly depicts humans as deeply self-interested.

One might protest against the conclusion reached by Frank et al. (1993), claiming that the most likely explanation is that individuals less likely to cooperate are more inclined to pursue a career in economics. To dismiss this hypothesis and support their view that the reduced cooperation is *caused* by training in economics (rather than merely *correlated* with it), Frank et al. (1993) compared students from economics and astronomy *before* they started their courses *and after* they completed them. Both economists and astronomers at the beginning of their courses presented similar degrees of cooperation. Students of economics were observed to reduce their cooperation much more than astronomers (Frank et al., 1993, p. 169). Corroborating these findings, Wang et al. (2011) also show that economics majors are less concerned with fairness and have a more positive view of greed than others. As we can see, believing in psychological altruism or psychological egoism has both theoretical and practical relevant implications.

Here is a summary of the four parts of this thesis. Part I aims to clarify the terminology used in the debate on psychological altruism and explain the philosophical context in which this debate is situated. In *Chapter 2*, I address the "many faces of altruism", explaining and comparing the main accounts of altruism in the scientific literature. I will give particular attention to the standard account of altruistic motivation, discussing it in detail. I also present alternative accounts of altruistic motivation and explain how they can be more theoretically useful than the standard account. Finally, I present an overview of the debate opposing psychological altruism and psychological egoism. In *Chapter 3,* I discuss key philosophical issues that are important for the debate on psychological altruism. I will address the main accounts of mental states in the tradition of philosophy of mind, as well as the different theories of desire, discussing their significance to the debate on psychological altruism. I highlight

several assumptions that underlie the debate on psychological altruism, arguing that some of these assumptions are problematic. This first part of the thesis aims to sets the stage for the discussions of the next parts.

Part II analyzes and discusses the scientifically oriented arguments for psychological altruism, which have shaped the contemporary debate. I will focus on the two most influential arguments of this debate. Firstly, in *Chapter 4*, I will discuss Daniel Batson's (1991, 2011) empirical case for the existence of altruistic motivation. Batson is a social psychologist who has dedicated much of his career to pursuing the answer to the question of whether altruism exists. He developed a set of ingenious experiments to test the motivation behind altruistic behaviors in humans. Batson believes that these experiments provide evidence for the truth of psychological altruism. But although Batson offered progress to the debate, I will argue that his theoretical framework has some flaws and point out some limitations of his empirical approach.

In *Chapter 5*, I discuss the *evolutionary* case for psychological altruism. There, I will focus on the most influential evolutionary argument, proposed by Elliot Sober and David Sloan Wilson in their book *Unto Others* (1998). I will argue that the evolutionary arguments for psychological altruism are also unconvincing. My conclusion in Part II is that the main scientifically oriented arguments analyzed fail. More importantly, I claim that they fail due to the fact that the concepts of altruistic and egoistic motivation, as well as the hypotheses of psychological altruism and psychological egoism, are all problematic and outdated ways of conceptualizing altruistic motivation. The standard account makes altruistic motivation an impractical, virtually unverifiable phenomenon. This, in turn, makes the standard account an unfruitful notion for scientific research.

Part III shifts the focus from contemporary science to the history of philosophy. It is commonly assumed that the debate opposing psychological altruism and psychological egoism is part of a long-standing debate, encompassing a great part of the history of philosophy, especially modern philosophy. This historical background gives legitimacy to the contemporary debate: when engaging in the debate over the existence of psychological altruism, authors can assume that they are taking part in something truly important in the history of ideas. However, I will discuss a series of problems with this historical assumption.

C*hapter 6* is dedicated to discussing the history of egoism in modern philosophy. I argue that authors often referred to as proponents of psychological egoism are not necessarily holding such a view. There are alternative interpretations of their egoistic views of human motivation, which do not entail or assume psychological egoism. More importantly, psychological egoism is not a central element of their philosophies as it is often assumed. *Chapter 7* is dedicated to the history of altruism. I argue that the authors who have put forward a less egoistic account of human nature — including Auguste Comte — were concerned with issues that are not captured by the idea of psychological altruism. They conceived altruistic motivation differently. The goal of Part III is to show that reading the authors analyzed through the lens of psychological altruism and psychological egoism oversimplifies and impoverishes their views. I argue that there is a disconnection between the historical debate on altruism and the contemporary approach. This helps to undermine the historical legitimacy of the contemporary debate. If this historical foundation is questioned, then we have more reasons to question the contemporary understanding of altruism and egoism.

Part IV focuses on the relationship between the standard account and the ordinary, non-technical use of the term “altruism”, discusses the normative dimension of altruism, and suggests

how we can conceptualize altruism in a way that can avoid the problems raised throughout this thesis. In *Chapter 8*, I will propose the term "*ordinary altruism*" to refer to the meaning of altruism as it is conceived in everyday, ordinary language. Differently from the technical account of altruism, I will not provide a definition with necessary and sufficient conditions for ordinary altruism, because this account does not fit into such a restrictive definition. Instead, I will investigate ordinary altruism through the discussion of prototypical cases of altruism. This investigation will support the main claim of Chapter 8, namely, that the standard account of altruistic motivation fails in representing ordinary altruism.

In *Chapter 9*, I introduce the issue of the normative character of altruism, pointing out how the standard account unjustifiably excludes it from altruism. I argue that there are good reasons for conceptualizing altruism as context-dependent. I also discuss some directions that may be adopted once we reject the standard account. I propose the idea of conceptualizing altruism as a *virtue* rather than a mental state. This is what I call *virtue altruism*. I claim that this account successfully represent ordinary altruism, accommodates its descriptive and normative dimensions, and is at home with most of the historical accounts of altruism. I also offer an argument to show that the standard account of altruistic motivation, on top of all its problems, also discourages altruistic actions. Beyond its theoretical benefits, virtue altruism can promote the active pursuit of altruistic actions.

Chapter 2

The Many Faces of Altruism

2.1 Altruism as a Behavior

There are two general tendencies in the interpretation of the meaning of "altruism" in the literature. One considers altruism to be a *behavior*, the other a *motivation*[2]. So, before using the term "altruism," one should specify which of the two interpretations one is adopting. But things get more complicated. If, for example, one wants to use "altruism" as a motivation, then one will face the challenge of determining *which* psychological states constitute altruistic motivation. A similar challenge occurs for the notion of altruism as a behavior: what are the conditions for a behavior to be considered altruistic? The literature offers a plurality of accounts of altruism, and each of these considers different features to be essential to altruism. I will call these the *technical accounts of altruism*. The goal of this chapter is to explain and compare the technical accounts of altruism in the literature, thus making sense of the many faces of altruism.

Firstly, I should state that the existence of divergent accounts of altruism is not a problem in itself. Different research projects have different goals, and the best way of characterizing altruism will vary together with these different goals. Certain accounts of altruism will be more useful than others, depending on the phenomenon of interest for researchers, so a rich conceptual repertoire can be helpful. The plurality of ways of conceptualizing altruism only becomes a problem when authors are not precise about which notion of altruism they are using, thus

[2] Another way of using the term "altruism" is to apply it to certain individuals, referring to stable personality traits. I will address this use in Chapter 8.

creating much confusion. Unfortunately, as Clavien & Chapuisat (2013) note, this confusion is common in the literature (p. 126). Philosophers can help, not by identifying the "right" account of altruism, but by clarifying the different accounts and explaining the implications of each of them. This is the approach I adopt here.

Here is the structure of this chapter. In this first section, I address the idea of altruism as a *behavior*, considering some of the challenges that these accounts face and their limitations. *Section 2.2* addresses the *evolutionary* account of altruism, which defines altruism in terms of fitness effects. These two initial sections address technical accounts of altruism that disregard the motivation underlying helping behaviors. Since the main object of study in this thesis is the *standard account of altruistic motivation*, this chapter gives particular attention to this account of altruism. In *Section 2.3*, I explain the standard account of altruistic motivation in detail. *Section 2.4* discusses some alternative accounts of altruism as a motivation, which compete with the standard account. Finally, in *Section 2.5* I discuss the debate opposing the hypotheses of psychological altruism and psychological egoism.

So, what makes a behavior "altruistic"? Clavien and Chapuisat (2013) offer a technical account of altruism, which they call "behavioral altruism", that aims to represent the way in which altruistic behaviors are usually conceived in the literature. The authors propose that a behavior is a case of behavioral altruism "if it brings any kind of benefit to other individuals at some cost for the agent, and if there is no foreseeable way for the agent to reap compensatory benefits from her behaviour" (Clavien and Chapuisat, 2013, p. 131). Notice that this account is all about the *effects* of the behavior, not about the agents' intentions. What matters, here, is whether the behaviors are *actually* costly and whether there is no foreseeable way for the agent to be compensated.

The notion of behavioral altruism seems to capture an intuitive view of what an altruistic behavior is: behavioral altruists are helping others, have some costs in doing so, and do not anticipate getting benefits for this in the future. However, we should also point out some challenges that researchers would face when using this account of altruistic behavior. Firstly, it demands a calculation of costs and benefits. But such a calculation is hard to obtain. Empirical studies often use monetary gains and costs as an indicator of costs and benefits, but this would not be a good method to analyze costs and benefits for an infant, for example. Furthermore, for us to know whether a behavior qualifies as a case of behavioral altruism, we should also determine whether there is a foreseeable way for the agent to receive benefits. Since we are talking about just any sort of benefit, it is hard to track when benefits are foreseeable or not. These conditions are demanding and particularly hard to apply in certain research programs.

Ramsey (2016) proposes a different technical account of altruism as a behavior, which he calls "*helping altruism*". This account is not so demanding as behavioral altruism, requiring neither a cost for the performer nor a proper evaluation of costs and benefits. The conditions for an action to be a case of helping altruism is simply that *it helps another individual* and that the helping is "a behavior that is not a mistake" (Ramsey, 2016, p. 35). The condition of not being a mistake aims to overcome a limitation in Clavien and Chapuisat's (2013) behavioral altruism. As Ramsey (2016) points out, if an "Olympic gymnast tragically falls down at the end of her routine, she benefits her competitors", but, Ramsey claims, "this is no altruistic act on her part" (p. 35). The gymnast's behavior is costly, benefits others, and there is no foreseeable way in which she will be benefited from this behavior. Helping altruism is supposed to exclude this kind of case from altruism.

Ramsey's helping altruism offers an account of altruism as a behavior that is not as demanding as Clavien and Chapuisat's behavioral altruism. These are useful concepts, and certainly different researchers, in different areas of study, will have different preferences. For example, helping altruism seems to fit better into Warneken & Tomasello's (2008) study of helping behaviors in infants (Ramsey, 2016, p. 34). My goal in this chapter is not to discuss which account of altruism as a behavior is the best, but merely present a repertoire of accounts, highlighting some of their advantages and disadvantages.

Regarding the challenges that one might face when using Ramsey's (2016) helping altruism, we should consider that the condition of "not being a mistake" is not established in a very clear way, leaving space for some degree of vagueness in this account. Ramsey (2016) suggests that one alternative to determine whether a behavior is a mistake is to ask whether it is an adaptation (p. 35). But Ramsey presents this as a suggestion, not integrating this as a condition of helping altruism. A second possible challenge for researchers interested in using helping altruism is that this account of altruism is quite inclusive. How can we make sense of the difference between mere "helping" and "altruism"? One can simply neglect this distinction, saying that both are ultimately the same, but it seems that something is missing here. Making altruistic behaviors such a broad category risks deflating the meaning of altruism altogether.

A third different account of altruism as a behavior is proposed by the primatologist Frans de Waal (2008). This account, called "*directed altruism*", describes a "helping or comforting behavior directed at an individual in need, pain, or distress" (de Waal, 2008, p. 281). Directed altruism can be produced spontaneously, through learning, and through the prediction of future consequences. As a primatologist, de Waal is interested in an account of altruism that could be

applied to non-human species, and directed altruism offers an account of altruism that could be used in different species.

Even though directed altruism is still a behavioral account of altruism, it requires the cognitive capacity to perceive others in need. In directed altruism, the proximate mechanism responsible for detecting when individuals are in need is *empathy*, defined broadly as sensitivity to others (de Waal, 2008, p. 282). As we will see in Chapter 4, empathy is considered to be the main basis of altruistic *motivation*., but is often neglected in the behavioral accounts of altruism. Rather than requiring costs for the agents, de Waal's directed altruism requires instead that agents feel empathy for others in need.

A positive aspect of directed altruism is that it seems, in principle, capable of avoiding some problems raised against the previously discussed accounts. Since it does not require the measurement of the agent's costs, directed altruism avoids the challenges raised against Clavien and Chapuisat's (2013) behavioral altruism. By requiring empathy, it can also avoid the overly inclusive character of the account proposed by Ramsey (2016).

However, directed altruism also faces some challenges. The scope of directed altruism seems to be too narrow. It restricts altruistic behaviors to cases in which the agents perceive others to be in need. But it also seems to make sense to talk about altruistic behaviors directed towards others who are not particularly in need. One might help others to make them better than they are, even if one believes that the receiver is already doing well. Also, de Waal (2008) claims that, "[b]y definition, altruism carries an initial cost" (p. 281). However, he is not clear about what these costs are and how we should measure them. Furthermore, de Waal's directed altruism demands a set of complex cognitive processes, which are required to perceive when others are in need. This might impose a set of requirements too restrictive for some researchers interested only

in helping behaviors. For example, an economist interested in the social variables increasing the presence of helping behaviors is likely to ignore the emotional underpinnings of the particular instances of helping behavior.

The accounts discussed here are relevant to many areas of research. But one area in which the phenomenon of altruistic behaviors is particularly important and challenging is economics. The idea of humans helping others without receiving a reward often conflicts with some views in this area. An influential model in economics, known as "*homo economicus*", represents humans as rational agents, calculating the cost and benefit of every action, guided by a monetary conception of pleasure. This model "reduces human preferences to individuals' own welfare — or payoff — maximization" (Clavien & Chapuisat, 2016, p. 25). If the model of the *homo economicus* is correct, we might have difficulties in explaining the existence of altruistic behaviors. The awareness that certain behavior is costly and will not pay off in the future (as in behavioral altruism) should preclude one from performing it. In this model, altruistic behaviors are more than uncommon or strange — they are *irrational*.

The notion of "utility" and the question of how to measure it are fundamental issues underlying the conflict between the *homo economicus* and altruistic behaviors. The reason why some altruistic behaviors seem to be in conflict with the idea of rational agents is that the utility function attributed to agents in the *homo economicus* model, at least in its simpler expression, excludes the possibility of including others' welfare in the agent's utility calculus.

Many authors have criticized the model of the *homo economicus*. The model is incompatible with the data from experimental economics, which shows humans consistently presenting behavioral altruism (Fehr & Fischbacher, 2004; Fehr & Gächter, 2002). Furthermore, the model assumes an unreal degree of rationality that seems incompatible with ordinary human

beings (Thaler & Sunstein, 2008, p. 6). More recently, researchers have proposed different models in which concern for others is included in the utility functions (see Clavien & Chapuisat, 2016). Since the *homo economicus* model is merely one of the many models available, there is no *a priori* reason to assume that altruistic behaviors are irrational[3].

As I said, it is not my goal here to determine which account of altruism as a behavior is the best one. The purpose of this section is to clarify terminology, showing the positive and negative aspects of each account. This discussion highlights how thinking about altruism as a behavior is difficult. But it should also be clear that it is useful: if researchers want to talk about altruistic *behaviors*, using a motivational account of altruism would only make things more complicated, imposing the need to make unnecessary assumptions about individuals' mental states. The behavioral accounts of altruism offer conceptual alternatives for these researchers. In the next section, I address yet another way of thinking about altruism. This is the evolutionary approach to altruism, which conceives altruism in terms of fitness effects.

2.2 Evolutionary Altruism and The Selection of Unselfish Phenotypes

The term "altruism" is part of the vocabulary of evolutionary biology since the work of Herbert Spencer (1862/1883). In this area, "altruism" has been used to refer to a subset of helping behaviors, namely, those helping behaviors that have an impact on the *fitness* of agents and recipients. This section explains the account of altruism that is mostly used in contemporary

[3] Peart and Levy (2005) argue that the difficulty in accommodating individuals' altruistic behaviors into the agent's utility calculus is a problem typical of the neoclassical economics. The authors claim that this was not so much of a problem for classical economists, for they accepted a "sympathetic principle", which includes in one's utility function the welfare of others (Peart & Levy, 2005, p. 191).

evolutionary biology, known as "*evolutionary altruism*". From all the different technical accounts of altruism, evolutionary altruism is probably the most well-established account of altruism in the scientific literature, figuring at the center of important debates in evolutionary biology. I also discuss the notion of *reciprocal altruism*, which is also an important account of altruism in evolutionary biology.

Evolutionary altruism (sometimes called "biological altruism") is defined as a behavior (or phenotype[4]) that benefits others and is costly to the performer, where costs and benefits are measured in terms of *direct fitness*[5] (West et al., 2007, p. 416). That is, evolutionary altruism describes traits that increase the fitness of other individuals at the expense of the agent's own direct fitness. Importantly, direct fitness's costs and benefits are measured in terms of one's *lifetime* fitness. Thus, even if an individual performs a series of costly behaviors that increase the fitness of others, it will not be evolutionarily altruistic if these end up paying off for that individual in the long run.

The existence of evolutionary altruistic traits poses a challenge to evolutionary biology. Natural selection occurs in populations with heritable phenotypic variation, where this variation has different rates of survival and reproduction in different environments (Lewontin, 1970).

[4] Although it is common for authors to define evolutionary altruism as a "behavior", calling it a "phenotype" is more precise. Phenotypes encompass behaviors but are not limited to them. Consider, for example, the case of the animals, such as guinea pigs and Himalayan mouses, that manifest their fur coloration differently depending on the temperature of their skin (Kidson & Fabian, 1981). The phenotype "skin color", in this case, may or may not be considered a behavior. By using the term "phenotype", we avoid the debate over what a behavior is.

[5] Direct fitness is defined as "the component of fitness gained through the impact of an individual's behaviour [or phenotype] on the production of offspring" (West et al., 2007, p. 416). If something increases the likelihood of having more offspring, then it increases one's direct fitness. Direct fitness is contrasted with indirect and inclusive fitness, which will be discussed below.

Individuals are expected to have traits that improve the likelihood of *their own* survival and reproduction. Says Darwin (1859/2008): "[t]his preservation of favourable variations and the rejection of injurious variations, I call Natural Selection" (p. 63). Traits that "benefit" (that is, increase the fitness of) their own bearers are more easily explainable in evolutionary terms. Using an agential language[6], we can say that natural selection privileges the most "selfish" individuals, that is, the individuals that aim to increase their own fitness (Dawkins, 1976/2006). This seems to suggest that any phenotype evolutionarily altruistic will be eliminated.

Nevertheless, cooperation is pervasive in nature. As an example, consider the phenomenon of eusociality. In this form of social organization, some individuals behave in ways that seem to be the opposite of what natural selection promotes: they take care of others' offspring, do not reproduce themselves, and even sacrifice themselves to protect the colony. Commenting on the existence of neuters or sterile females in insect communities, Darwin (1859/2008) said: "[it] at first appeared to me insuperable, and actually fatal to my whole theory" (p. 175). The question that is raised here is: how could these altruistic phenotypes evolve? Why has not natural selection eliminated such selfless traits? This problem is known as "the problem of altruism" and has produced much discussion throughout the history of evolutionary biology (Carter, 2005).

With the development of population genetics, beginning in the 1930s with the work of J. B. S. Haldane, S Wright, and R. Fisher, evolutionary biologists finally had the theoretical instruments to explain the evolution of altruistic phenotypes. An infamous attempt to answer the problem of altruism is that of the British zoologist, Vero Copner Wynne-Edwards. In his book

[6] The use of agential language, representing individuals/genes/groups as agents with goals, is a useful and widely used resource in evolutionary biology (see Okasha, 2018).

Animal Dispersion in Relation to Social Behavior (1962), Wynne-Edwards defended a theory called "group selection". In this theory, rather than considering *organisms* as the units of selection, Wynne-Edwards consider *populations* as the units subjected to natural selection. He believed that many traits in nature could only be explained as traits selected for the benefit of the species (see Borrello, 2010). For example, he thought that populations could regulate their density, and that this could not be explained by individual selection. If group selection were correct, the selection of evolutionary altruism could be explained as a *group-level adaptation*: although these altruistic traits are costly for the individual, they may be selected for their benefit to the group. Wynne-Edwards's theory was heavily rejected by other evolutionary biologists (e.g., Williams, 1966/2018; Dawkins, 1976/2006). After the rejection of Wynne-Edwards's theory, the term "group selection" became "a code word for a sloppy, mushy, and confused view of how natural selection operates"[7] (Wilson, 2005, p. 172).

Shortly after Wynne-Edwards's publication of his defense of group selection, the biologist William Hamilton proposed what became the most influential solution to the problem of altruism. Hamilton (1964a, 1964b) argues that in the same way that a phenotype can be selected for increasing the organism's number of offspring, it can also be selected for increasing the number of offspring produced by one's *relatives*. In his theory, Hamilton distinguishes *direct* from *indirect* fitness. Direct fitness is measured on the basis of the number of offspring an organism produces, while indirect fitness is the component of fitness obtained from benefiting

[7] However, group selection has received a renewed attention recently. A notorious proponent of what is known as "new" group selection theory is David Sloan Wilson, who has defended a reformed account of group selection since the 1970s (Wilson, 1975). In his book coauthored by E. Sober, *Unto Others* (1998), Wilson presents his multilevel selection theory, arguing that the selection of groups is more than a theoretic possibility, but something supported by empirical evidence. Among the authors favoring a pluralistic account of natural selection that includes groups as units of selection is the father of sociobiology, E. O. Wilson (Nowak et al., 2010).

related individuals (West et al., 2007, p. 416). Putting both inclusive and direct fitness together, we have what Hamilton called "*inclusive fitness*". With the notion of inclusive fitness, Hamilton provided an elegant solution to the problem of altruism. Altruistic phenotypes can be costly in terms of direct fitness, but they can still be selected due to their indirect fitness effects. Indirect fitness can compensate for the costs of direct fitness, thus explaining how altruistic phenotypes can evolve.

In Hamilton's approach, instead of taking *organisms* as the *units of selection*, we consider *genes* as the proper units under selection. As popularized by Richard Dawkins (1976/2006), organisms are mere "vehicles" for our "selfish" genes. By benefiting their relatives, organisms are indirectly promoting the interest of their own genes. The sterile individuals in eusocial species, for example, who take care of others' offspring, can evolve because they share genes with these individuals that they are helping. Of course, whether or not an altruistic phenotype is selected depends on how costly it is, how much it benefits others, and the degree of relatedness between the agent and the beneficiaries. Taking these three variables into account, Hamilton's theory can make predictions about whether altruistic phenotypes can be selected.

Inclusive fitness theory is still dominant today, finding strong empirical support (Abbot et al., 2011). However, it does not explain all cases of helping behaviors in nature. The notion of *reciprocal altruism*, proposed by Robert Trivers (1971), is another account of altruism in evolutionary biology. Reciprocal altruism describes behavior that benefits a recipient but is compensated by the recipient in future interactions. Reciprocal altruism is particularly important due to its potential to explain cooperation between different species, which is a phenomenon that cannot be explained by Hamilton's inclusive fitness theory. To illustrate this, consider the

example of the plover bird's cleaning symbiosis with crocodiles, where crocodiles leave their mouth open while birds eat the food between their teeth. How could this behavior evolve?

We can sum up the selection of reciprocal altruism as follows: if (1) an agent has a behavior B1 that benefits a recipient, and (2) behavior B1 causes the recipient to perform a behavior B2 that benefits the performer, where (3) B2 *outweighs* the cost of B1, then (4) B1 can be selected. Applying this to the example above, we can say that if (1a) the crocodile benefits the bird by not eating it and offering free food in its teeth, and (2a) the cleaning of the crocodile's teeth is beneficial for the crocodile, and (3a) the benefits of the cleaning outweigh the costs of not eating the bird, then (4a) this behavior can be selected. The crocodile "invests" by not eating the bird, and this is more advantageous in the long run[8].

Although reciprocal altruism became a very popular term in evolutionary biology, many have resisted classifying it as "altruistic". As I said, evolutionary altruism became a fundamental term in evolutionary biology. So, many authors in biology call evolutionary altruism simply "altruism", assuming that this is what the word "altruism" means in the context of evolutionary biology. This makes biologists reject reciprocal altruism as an account that is "really altruistic". They claim that "reciprocal altruism is not altruistic" for "it provides a direct fitness advantage to cooperating" (West et al., 2007, p. 420). Thus, to avoid the confusion of using "altruism" in different senses, some authors prefer to use the term "reciprocation" instead of "reciprocal altruism" (West et al., 2007, p. 417). Regardless of whether reciprocal altruism is an appropriate term or not, what is important is to keep in mind that it is fundamentally different from evolutionary altruism.

[8] Although particularly helpful to explain helping behaviors between different species, reciprocal altruism is also widespread in within-species relations. A famous example is the behavior of sharing regurgitated blood among vampire bats (Wilkinson, 1984).

The first two sections of this chapter discussed many accounts of altruism that disregard the motivation underlying the helping behavior. These accounts of altruism form a complex repertoire, allowing us to use the term "altruism" in many ways. Now, I move on to the approach to altruism that conceives it to be a *motivation*. The next section is dedicated to the standard account of altruistic motivation.

2.3 The Standard Account of Altruistic Motivation

Purely behavioral definitions of altruism are criticized for neglecting motivations, which many consider to be a fundamental aspect of altruism (Peacock et al., 2005). The most common view of altruism describes it as a form of motivation (Doris et al., 2020). But altruistic motivation can be defined in different ways. In the standard account of altruistic motivation, altruistic motivation is defined as an *ultimate desire to increase the welfare of others* (Batson, 2011; Sober & Wilson, 1998). So, when discussing altruistic motivation, we are not talking about just *any* motivational state to help others, but a very specific mental state. In order to understand altruistic motivation, we need to understand what is meant by "desire", "ultimate", and "welfare". This section aims to explain these elements.

Desires, just like beliefs, are mental states that represent certain states of affairs. We say of these mental states that they are *intentional* or that they *have intentionality*[9], that is, that they are *about* certain states of affairs. Although both are about states of the world, they have different

[9] Intentionality is a technical term in philosophy that can easily be misinterpreted. The meaning of "intentionality" should not be confused with the meaning of "intention". Our intentional states are not necessarily voluntary or planned. They are simply *about* certain states of affairs. Notice also that beliefs and desires are the main intentional states, but other states, such as hopes and fears, are also intentional.

directions of fit (Searle, 1983, p. 7). Beliefs aim to represent the world *as it is*. We believe in a proposition when it seems to match the world. If it actually represents the world, it is a true belief. By contrast, desires represent *what we want* the world to be. Our desires do not aim to match the world but to represent the states of affairs we want to be true. In short, the direction of fit of beliefs is mind-to-world (the mind adapts to the world), while that of desires is world-to-mind (the world adapts to the mind).

The interaction between our desires and our beliefs in decision-making makes the basis of our deliberate interactions with the world (Stich et al., 2010, p. 150). Actions are guided by what we want (desire) and by the best way to bring it about, given what we think the world is (beliefs). This way of understanding the interaction between beliefs and desires is, to some degree, present in philosophy at least since Aristotle (Kahn, 1987). Aristotle (2000) says that "[p]ursuit and avoidance in the sphere of desire correspond to affirmation and denial in that of thought" (p. 104). The interaction between beliefs and desires is summarized by Hobbes (1651/1998) in an ingenious metaphor: "the thoughts, are to the desires, as scouts, and spies, to range abroad, and find the way to the things desired" (p. 48).

In contemporary philosophy, the fundamental feature of intentional states is that they have *contents*. The content of an intentional state is a representation of the world, determining what this state is *about* (Fodor, 1981, p. 25). The widely accepted view, both in philosophy of mind (Schroeder, 2020) and in the debate on psychological altruism (Sober & Wilson, 1998, p. 212), is that the representational content of beliefs and desires are *propositions*. Propositions are claims about the world, which are either true or false. To believe in something is to take a given proposition to be true, and to desire something is to wish a proposition to be true. Since beliefs and desires share the same kind of content, they can interact in decision-making.

The technical account of desire that has been discussed here should be distinguished from the colloquial meaning of desire. In everyday language, we often restrict the term "desire" to describe cases where one wants something in a particularly strong or passionate way (Schroeder, 2004, p. 4; Gregory, 2021, p. 6). By contrast, the technical philosophical notion of desire encompasses what we, in common parlance, would describe using different words, such as wants, intentions, cravings, wishes, and so on. Desire, as I am discussing here, is not necessarily strong or passionate, thus departing to some degree from its meaning in ordinary English[10] (see Searle, 1983, p. 29; see also Arpaly & Schroeder, 2014, p. 115).

In order to understand the notion of altruistic motivation, the key distinction that should be clear is that between *instrumental* and *ultimate* desires. This reflects a long-established distinction in philosophy between things that we want only as a means to something else and things that we want for their own sake, which was already discussed in Plato (e.g., 1997, p. 998) and Aristotle (e.g., 2000, p. 4). A desire is instrumental when it is only a *means* to the realization of another desire. By contrast, a desire is *ultimate*[11] when we want the desired state *for its own sake*, regardless of whether it would lead to the satisfaction of another desire (Arpaly & Schroeder, 2014, p. 6). For genuine altruistic motivation, the desire to help others needs to be ultimate. The debate on psychological altruism is centered on the question of whether our desires to help others can be ultimate or whether they are always instrumental to some egoistic deeper desires (Sober & Wilson, 1998, p. 201).

[10] In order to avoid confusion between the everyday and the technical accounts of desire, some authors prefer to use alternative terms, such as "conative states" (e.g., Schulz, 2018). However, I will use the term desire, since this is the term that appears in the definition of psychological altruism and is the term that most of the literature on psychological altruism adopts.

[11] Ultimate desires are also called "primary desires" (Clavien, 2012) and "intrinsic desires" (Schroeder, 2004).

Instrumental desires are conditional on the desires motivating them. But instrumental desires are also conditional on the *beliefs* based on which they were formed. To illustrate this, consider an example. Assume that I have an ultimate desire for pleasure. Assume also that I *believe* that playing guitar causes me pleasure. With this desire and this belief, I can create an instrumental desire to play guitar. This instrumental desire is conditional on *both* the ultimate desire to feel pleasure *and* the belief that playing guitar causes pleasure. If one of them is missing, the instrumental desire is gone. Desires and beliefs interact in an inferential process in our decision-making, where we create auxiliary desires and rely on our existing beliefs in order to get what we desire ultimately.

Although desires have a primary role in motivation, it is not the case that we will always act in ways to bring about what we desire. Individuals might have altruistic motivation and not produce any helping behavior. One of the reasons for this is that different desires can *conflict* with each other. The realization of one desire can contradict the realization of another. My desire to play guitar might conflict with my desire to read more. When this happens, we need to consider the different *intensities* of desires. It is often the case that the *stronger* desire will overcome the weaker, taking primacy in the production of behaviors (Schroeder, 2004, p. 13). Interestingly, the weaker desires that fail to motivate actions will not necessarily disappear. Desires might persist even if they fail to produce behavioral outputs. Since the definition of altruistic motivation refers to the ultimate desire itself, an ultimate altruistic desire that does not produce behaviors also qualifies as altruistic motivation. The existence of desires that do not produce behaviors is a phenomenon of particular interest for us to understand the debate on psychological altruism.

Individuals might have genuine altruistic motivation and not indicate that in their behaviors. The main challenge for the debate on psychological altruism, however, is that individuals might have ultimate altruistic desires *without knowing* that they have them. This is so because some desires are *unconscious*. Many desires indirectly influence our behavior, and we do not have access to all of them (Goldman, 1970, p. 122). The causal chain linking desires and beliefs often escapes from our conscious access (Sober & Wilson, 1998, p. 217). Individuals might have ultimate altruistic desires without behaving altruistically and without having access to these desires. A more challenging possibility, however, is that individuals can desire one thing, and believe that it is an ultimate desire, while, in reality, it is merely instrumental to an unconscious desire. That is, we might help others moved by a desire to help, which seems to us to be an ultimate desire, while this desire is actually instrumental to an ultimate egoistic desire. This scenario represents the main challenge in the debate on psychological altruism, and I will discuss it in more detail in the next chapter.

Another issue that can also cause confusion is the possibility of *mixed motives*. One might expect that psychological altruism says that some actions are based *solely* on altruistic desires. However, this is not the case. The hypothesis only demands the existence of ultimate altruistic desires, allowing mixed motivation. That is, a given action might be motivated by *both* egoistic *and* altruistic desires and still make psychological altruism true. This might sound counterintuitive. For example, imagine a politician who constructs a hospital in a city motivated by a desire to be reelected. If this politician also has an altruistic desire to help these people, even if this desire, alone, would not motivate him to do so, he still is considered to have genuine altruistic motivation. The standard account is permissive in this regard.

The final element of the definition of altruistic motivation that needs to be explained is the notion of "*welfare*". As the debate focuses on the notion of desire, the notion of welfare is often left in the margins, being interpreted differently by different authors. Welfare is, broadly speaking, that which is ultimately valued[12]. A desire to increase the welfare of others is a desire for the good of others. But what, exactly, constitutes welfare and wellbeing is something disputed. A hedonistic perspective might want to reduce welfare to one's hedonistic states; an Aristotelian ethicist might claim that virtues are a crucial part of welfare; and so on. Authors in the debate on psychological altruism often avoid this debate, leaving open the meaning of welfare (e.g., Schulz, 2016, p. 16).

The philosopher Philip Kitcher (1993, 1998, 2010, 2011) offers a discussion of the notion of "welfare of others" in altruistic motivation. As he explains, there are different ways of conceiving the welfare of others. On the one hand, the agent might desire to do what the *agent* believes will increase others' welfare. On the other hand, an agent might desire to do what the *beneficiaries* themselves believe will increase their own welfare. Kitcher proposes the term *paternalistic altruism* for the former and *non-paternalistic altruism* for the latter. In the literature, it is common for authors to adopt a *paternalistic* point of view. I return to this in Chapter 9.

This section introduced the standard account of altruistic motivation, explaining the way in which this account is understood in the literature. In the next chapter, I address the theoretical framework involved in the notion of altruistic motivation, discussing different views on mental

[12] Some authors prefer using the term "wellbeing" rather than "welfare" (e.g., Schulz, 2016; Piccinini & Schulz, 2019). "Wellbeing" might be a better term, but I will use "welfare" to remain consistent with the majority of the literature. I take these two terms as interchangeable, here.

states and the main theories of desire. In the next section, I present different motivational accounts of altruism.

2.4 Alternative Accounts of Altruistic Motivation

Although the standard account defines altruistic motivation as an ultimate desire, there are alternative motivational accounts of altruism. In this section, I will present two alternative accounts. Firstly, I will present the notion of *preference altruism*, proposed by Clavien and Chapuisat (2013). I then present *Kitcher's altruistic motivation*, proposed by Kitcher (2010, 2011). These accounts show less restrictive ways of thinking about altruistic motivation[13]. In this section, I will also discuss three cases in which altruistic motivation plays an important role: altruistic punishment; helping behaviors in infants; and organ donation. I argue that the alternative accounts of altruism presented here are likely to do a better job than the standard account in the context of the cases discussed.

Clavien and Chapuisat (2013) noted that, in many cases, researchers use the term "altruism" to refer to behaviors that are performed with the *aim* of benefiting others, regardless of whether this is merely instrumental for the agent to receive further benefits. In this case,

[13] Less restrictive accounts of altruistic motivation could also be relevant for discussing motivation in other species. In mentioning this, I should acknowledge the fact that, although my discussion of altruistic motivation is limited to humans, I do not reject that other animals might have altruistic motivation. If psychological altruism is grounded in human empathy/sympathy mechanisms, as Batson (2011) argues, and if these mechanisms are shared by other animals, as de Waal (2012) argues, then these non-human animals might have altruistic motivation (see also de Waal, 2006). However, as altruistic motivation is a higher-order motivation, it demands a set of complex cognitive skills. There is a set of issues in cognitive ethology that would have to be addressed before applying altruistic motivation to other animals (see Allen & Bekoff, 1999). So, in order to avoid all these problems, this thesis discusses altruistic motivation only in humans.

neither their accounts of behavioral altruism not the account of evolutionary altruism can be used, for they do not require agents to have any kind of motivation. It seems that the standard account of altruistic motivation would be the best alternative, but this account is also inadequate. This is because the standard account of altruistic motivation rules out cases in which helping is *instrumental*. Clavien and Chapuisat (2013), then, claim that we need a different account of altruism to account for this particular use: a motivational account without the strong conditions imposed by the standard account of altruistic motivation.

To account for a more permissive use of altruistic motivation, Clavien and Chapuisat (2013) propose a new account of altruism, which they call "preference altruism". They define preference altruism as a behavior[14] that was caused by "preferences for improving others' interests and welfare at some cost to oneself" (Clavien & Chapuisat, 2013, p. 131). This motivation, however, can be instrumental to egoistic ultimate desires. Preference altruism resembles the standard account of altruistic motivation, however, it is "more explicit about the cost and less restrictive regarding the underlying psychological mechanism" (Clavien & Chapuisat, 2013, p. 131).

The term "preference" in "preference altruism" aims to be a broad way of talking about motivation. It aims to make clear that, in preference altruism, we are not talking about ultimate desires. Any sort of willingness to help others is sufficient. The use of "preference" in "preference altruism" should be distinguished from the *relational* use of preference (see

[14] Notice that, although this account is still about behaviors, it has a motivational component. Some authors also refer to the standard account of altruistic motivation in a similar way. They say, for example, that "[a] behavior or an action is psychologically altruistic if and only if it is motivated by an ultimate desire for the well-being of others" (Stich, 2016, p. 4). But we should understand that the standard account of altruistic motivation refers to the motivation itself (Sober & Wilson, 1998). The behavior motivated by it might be also called "altruistic", but the fundamental phenomenon to which the definition refers to is the motivational state.

Hausman, 2012). In the latter use, we always prefer something in comparison to something else (see Gregory, 2021, p. 6). "If there are only two alternatives, one can *desire* both, but one cannot *prefer* both. Because they are comparative, preferences, unlike desires, require that one weigh alternatives" (Hausman, 2012, p. x). However, some authors use desires and preferences interchangeably (e.g., Sterelny, 2003), and this seems to be the use intended by Clavien and Chapuisat (2013).

A different way of defining altruistic motivation is presented by Kitcher (2010, 2011). Although he calls this "psychological altruism", I use the term "Kitcher's altruistic motivation" in order to avoid terminological confusion.

> A acts psychologically altruistically with respect to B in C just in case
>
> (1) A acts on the basis of a desire that is different from the desire that would have moved A to action in C*, the solitary counterpart of C (A).
> (2) The desire that moves A to action in C is more closely aligned with the wants A attributes to B in C than the desire that would have moved A to action in C*.
> (3) The desire that moves A to action in C results from A's perception of B's wants in C.
> (4) The desire that moves A to action in C is not caused by A's expectation that the action resulting from it would promote A's solitary desires (with respect to C and C*). (Kitcher, 2011, p. 22)

In short, conditions (1) to (3) state that the altruist agent desires to help a beneficiary after perceiving that the beneficiary wants something (i.e., is in need). As Kitcher (2011) explains, one could meet these first three conditions and still be moved by egoistic motivation (p. 23). The key element of altruistic motivation in Kitcher's definition is the condition (4). This last condition aims to exclude the possibility that the performer is benefiting the recipient only because she *expects* some personal benefit out of it. Kitcher (2011) calls the fourth condition the "anti-Machiavellian" condition (p. 23). Interestingly, Kitcher does not give an emphasis on the

distinction between ultimate and instrumental desires. His definition is focused on ruling out cases where the desire to help others follows from one's expectations of receiving rewards. Kitcher (2010) explains that this account aims make altruistic motivation compatible with formalizations from game theory and rational decision theory (p. 124).

Preference altruism offers a less restrictive way of talking about altruistic motivation than that of the standard account. Kitcher's altruistic motivation offers a more precise definition and also avoids the notion of "ultimate desire," which is the central problem in the standard account of altruistic motivation. Now, to see how these alternative accounts can be more theoretically useful, consider a few uses of altruistic motivation in the literature.

The first case I will address is the phenomenon known as "altruistic punishment", which occurs when agents punish free-riders, even though the punishment is costly for the agent (Fehr & Gächter, 2002). For researchers interested in variables such as the evolutionary conditions for the occurrence of altruistic punishment, or the social structures that make it more prevalent, behavioral altruism might be the best account of altruism. But this is not always the case. Researchers might be interested in the *motivational* components of altruistic punishment.

One of the controversies about experiments that identify altruistic punishment concerns whether the agents *want to benefit* the members of their group through the punishment of free-riders or whether what they really want is to *punish* the free-riders, benefiting the members of the group as a by-product (Clavien & Klein, 2010). Individuals in the former option want to help others. But this does not mean that they want it "ultimately". They might want to help others because they enjoy doing so, for example. So, the standard account of altruistic motivation would not be useful to make the distinction proposed here. Since we are talking merely of

individuals who want to help others, either ultimately or instrumentally, preference altruism would be the most suitable account.

Altruistic motivation is also used in the study of infants. The work of Warneken and Tomasello (2006) on infants is a famous example. They tested whether infants tend to help unknown adults perceived to be struggling to perform a certain task, such as putting an object into a cabinet with the doors closed. Their results show that toddlers consistently help, even though there is no obvious reward for them. In a subsequent study, with 20-month-old toddlers, Warneken and Tomasello (2008) have shown that, in some cases, the presence of rewards is not only unnecessary but can even *reduce* the helping behavior. When socializing practices involving rewards are in place, infants were less likely to help in comparison to cases where help was not associated with rewards (Warneken & Tomasello, 2008, p. 1787). This tendency to help is corroborated in studies such as Barragan and Dweck (2014), which shows that a simple social interaction suffices to trigger helping behaviors in 1-2 years-old toddlers, and Hamlin and Wynn (2011), which shows evidence that prosocial tendencies can be already observed even in 5 months-old infants.

It is common for studies such as the ones mentioned above to use the term "altruism". So, we can ask whether the phenomenon they are discussing could be properly represented by the standard account of altruistic motivation. Warneken and Tomasello (2006) talk of "altruistic motivation", but they refer to the "effort to help another person — with no immediate benefit to oneself" (p. 1301). The authors do not seem concerned with the idea that the helping behavior is ultimately based on an internal, subjective reward. In the introduction of Warneken and Tomasello (2008), they ask: "do human beings help one another because the helpful act itself is inherently rewarding or only because the helpful act is instrumental in bringing about separate

outcomes such as material rewards or the avoidance of punishment?" (p. 1785). The authors are concerned with ruling out helping behaviors motivated by one's expectations of external or social rewards. Their account of altruism seems to allow cases where, for example, an agent enjoys the feelings produced by seeing others doing well and helps as a means to feel these pleasant feelings. Thus, the standard account of altruistic motivation does not seem to be a proper representation of what these authors are calling "altruism".

Should Warneken and Tomasello (2006, 2008) use preference altruism instead? This does not seem to be the case. Preference altruism demands "costs" for the agent. But it is not clear to what degree the infants' helping behaviors are costly. What would count as a "cost" for an infant in the circumstances of these experiments? The researchers seem interested in ruling out external rewards, but the need for the behavior to be costly is not a central issue.

Kitcher's altruistic motivation focuses on ruling out cases where the desire to help others follows from one's expectations of receiving rewards. So, in my view, Kitcher's account offers Warneken and Tomasello (2006, 2008) a better account of altruistic motivation, for it focuses on excluding cases where individuals are helping moved by the expectation of rewards, moving away from the obscure language of instrumental and ultimate desires.

The final case where altruistic motivation is used that I will address in this section is the regulation of the donation of bodily materials, especially organs. This is a challenging bioethical problem where altruistic motivation plays a significant role. In this case, having an appropriate definition of altruistic motivation can have a direct impact on the life (and death) of patients. This case offers a good test for the standard account of altruistic motivation.

In determining who can donate organs, authorities are in a dilemma. On the one hand, they want to receive organs from living individuals, for this is often the only means to save some

patients. Considering this, it makes sense to stimulate potential donors by promoting policies that support them financially. However, on the other hand, financial compensation might encourage people in vulnerable socio-economic conditions to "sell" their organs, which is an ethically repugnant outcome. So, how can we, at the same time, encourage the life-saving practice of voluntary donation of organs and avoid the ethical risks it involves? Here is where the notion of altruistic motivation is used: an ethically acceptable donation of one's organ is often considered to be motivated by altruistic motivation.

As Moorlock et al. (2014) state, the idea that the donation of organs has to be altruistic is widely adopted as a principle for orienting policies that regulate organ donations. In Europe, donations of bodily materials are exclusively based on the assumption that donors need to have altruistic motivation (Pennings, 2015). As an example, Moorlock et al. (2014) discuss a report from The Nuffield Council on Bioethics[15] (2011), in the UK. In this report, it is said that "[a]ltruism, long promulgated as the only ethical basis for donation of bodily material, should continue to play a central role in ethical thinking in this field" (Nuffield Council on Bioethics, 2011, p. 5). Donation is expected to be altruistic in the sense of being a "selfless gift to others without expectation of remuneration" (Nuffield Council on Bioethics, 2011, p. 120). Moorlock et al. (2014) state that one of the claims used in rejecting organ donations is that they are not altruistic, and urge that "[s]ince the application of altruism results in the rejection of potentially life-saving (and life-enhancing) donations, a precise, rigorous and justifiable definition is required" (Moorlock et al., 2014, p. 134).

[15] This is an independent organization that discusses pressing bioethical issues, making recommendations to the government and engaging with media debates.

The Nuffield Council on Bioethics (2011) defines altruistic motivation as "concern for the welfare of the recipient of some beneficent behaviour, rather than by concern for the welfare of the person carrying out the action" (p. 139). This seems very similar to the standard account of altruism. But despite the similarity, they differ in a crucial way. The Nuffield Council on Bioethics (2011) makes sure to state clearly that the motivation to help can be ultimately based on subjective rewards, such as one's pleasure in helping (p. 139). One can donate moved ultimately by the belief that this will make them feel good and this will still be considered altruistic in their account[16]. It is not obvious which account of altruistic motivation would be the best account for this case. But what is important is to note that the standard account of altruistic motivation, which is the dominant account of altruism among philosophers, is certainly not the best concept to be used in this context. All three examples lead to the same conclusion: the standard account does not offer a useful concept for researchers interested in altruistic motivation.

2.5 The Debate on Psychological Altruism

The previous four sections defined and discussed different accounts of altruism, giving particular attention to the standard account of altruistic motivation. The present section is dedicated to discussing the debate opposing psychological altruism and psychological egoism. As mentioned in the introduction, the contemporary debate on altruistic motivation opposes two hypotheses, *psychological altruism*, which claims that *some* of our ultimate desires aim to

[16] The authors say that they "do not think it important from an *ethical perspective* [emphasis added] that altruism is thoroughly 'pure'" (The Nuffield Council on Bioethics, 2011, p. 139). This is important for it states that their choice of definition is guided by ethical reasons.

increase *others'* welfare[17], and *psychological egoism*, which claims that *all* of our ultimate desires aim to increase *our own* welfare. This section discusses some key structural aspects of this debate. A clear understanding of the logic underlying the debate will be helpful for the discussion, in the next chapters, of arguments for and against these hypotheses.

First, I need to make an important clarification. The term "psychological altruism" is used ambiguously in the literature (see Schefczyk & Peacock, 2010, p. 170). This term usually refers to a hypothesis (or a theory), as I presented it here. However, it is sometimes used to describe altruistic motivation itself. That is, psychological altruism (the hypothesis) states the existence of "psychological altruism" (the motivation). The same ambiguity occurs for psychological egoism. Authors often apply "psychological egoist" to someone motivated by an egoistic ultimate desire rather than to someone *exclusively* motivated by egoistic ultimate desires[18] (Sober & Wilson, 1998, p. 202; Piccinini & Schulz, 2018, p. 5; Piccinini & Schulz, 2019, p. 4). In order to avoid confusion, I will restrict the terms "psychological altruism" and "psychological egoism" to describe the *hypotheses* mentioned in the previous paragraph, and use "altruistic motivation" and "egoistic motivation" to describe *motivational states*. Although there

[17] It is common for authors to refer to altruistic and egoistic desires as "other-directed" and "self-directed" desires, respectively (Sober, 2013, p. 149). I avoid using such terminology due to its vagueness. The fact that a desire is other-directed (that is, that it has someone other than the agent as its object) does not entail that it is a desire to benefit the other. There are other-directed desires to harm others, for example.

[18] This ambiguity is particularly problematic for egoistic motivation. The existence of altruistic motivation *implies* the truth of the hypothesis of psychological altruism. But the existence of egoistic motivation does not imply the truth of the hypothesis of psychological egoism: one can have both egoistic and altruistic ultimate desires. Furthermore, even if the ultimate desires of some individuals are exclusively egoistic, there might still be other individuals with altruistic ultimate desires, which would suffice to establish the truth of psychological altruism.

are different motivational accounts of altruistic motivation, from now on in this thesis, when I mention “altruistic motivation”, I am referring to the standard account of altruistic motivation.

Another point I should mention is that, although authors often draw ethical conclusions from the debate on psychological altruism, this debate, in itself, is *purely descriptive*, not normative. Psychological altruism and psychological egoism describe certain psychological states. They are claims about the world, which can be either true or false. These hypotheses and the states they describe have no intrinsic moral value (see Sober & Wilson, 1998, p. 237).

To understand the debate on psychological altruism, we should make it clear what is logically entailed by the two hypotheses. Psychological altruism is an *existential* claim, stating that there are *some* altruistic ultimate desires. It says that, even though not every desire to help is ultimate, and not everyone has altruistic ultimate desires, at least some people, sometimes, desire to help others ultimately. By contrast, psychological egoism is a *universal* claim, saying that *every* ultimate desire or *every* human being is egoistic. Notice that there is a logical *asymmetry* between these two hypotheses. The existence of a single altruistic ultimate desire is *sufficient* to make psychological altruism true and to *falsify* psychological egoism. By contrast, the existence of an egoistic ultimate desire neither falsifies psychological altruism nor makes psychological egoism true. Therefore, the conditions to falsify each hypothesis are different.

The two hypotheses are incompatible. However, they are not contradictory, strictly speaking. They are only contraries. The crucial difference is that contraries allow both hypotheses to be false, while contradictions require one of the claims to be true. Let me explain. In contradictions, the truth of one claim implies the falsity of the other *and* the falsity of one claim implies the truth of the other. By contrast, in contraries, the truth of one claim implies the falsity of the other, but the falsity of one claim *do not* imply the falsity of the other. That is, both

contraries can be false at the same time. Remember that psychological egoism claims that every ultimate desire is egoistic. This leaves no room for altruistic ultimate desires. Thus, *at least one* of the hypotheses must be false. Hence, the truth of one implies the falsity of the other. However, the falsity of one *does not* imply the truth of the other. They can be both false.

To understand why the two hypotheses are contraries rather than contradictories, we need to understand that some desires might be *neither* egoistic nor altruistic (Schulz, 2016, p. 16). One might have, for example, a desire to follow moral rules for their own sake, regardless of their consequences to others or oneself. This desire does not involve the welfare of others nor that of the agent, so it is neither egoistic nor altruistic (Sober & Wilson, 1998, p. 237). The existence of at least one ultimate desire that is neither altruistic nor egoistic would suffice to falsify psychological egoism — but it would not imply the truth of psychological altruism. Thus, from the falsity of psychological egoism, we cannot infer the truth of psychological altruism and vice versa. The falsity of psychological egoism is a *necessary* (but not *sufficient*) condition for psychological altruism to be true[19].

As I said before, the debate is centered on the dispute about whether psychological altruism is true, that is, whether altruistic motivation exists. But how can we know whether a given motivation to help is a case of altruistic motivation? Two responses that one can intuitively give are to use introspection and to infer the motivation based on the observation of the agent's behaviors. Let me briefly explain the challenges for each of these two responses.

[19] Sober and Wilson (1998) propose the term "motivational pluralism", which they define as "the hypothesis that some of our ultimate desires are altruistic while others are egoistic" (p. 333). This term might encourage a common misconception. Motivational pluralism, as opposed to psychological egoism, *seems* to refer to the idea that humans have all sorts of ultimate desires, which is a view held by some authors (e.g., Schroeder, 2004). But Sober and Wilson's motivational pluralism refers exclusively to altruistic and egoistic ultimate desires. Since I believe that the term can be misleading, I will avoid using it here.

Intuitively, the simplest way of accessing our desires and figuring out whether they are ultimate is using *introspection*. However, although some of our desires can be accessed, there are others that are *unconscious* (Schroeder, 2004, p. 15). This means that we cannot dismiss the possibility that, for any given desire we have, this desire is instrumental to another desire to which we do not have conscious access. So, although philosophers have relied heavily on introspection throughout history, today there is widespread skepticism regarding an introspective solution to the altruism question (see Sober, 2013, p. 152).

The limitation of introspection in the debate on psychological altruism has pushed authors to search for a different way to address the issue. An alternative strategy is to look at the agents' *behaviors* and infer whether their desires to help are ultimate. This alternative is particularly appealing for researchers trying to answer the altruism question empirically. However, this approach also runs into some problems. Individuals with egoistic ultimate desires can have *instrumental* desires to help others, so they can behave in the same way that someone with an ultimate desire to help would behave. So, any inference based on agents' behaviors is epistemically risky.

Both introspection and behavioral analysis run into difficulties, leaving us without an easy answer to the altruism question. This conundrum is what Clavien (2012) calls the "deadlock problem", which is the central problem in the debate over psychological altruism. Philosophers and scientists have proposed different strategies to avoid the deadlock problem, trying to answer the altruism question despite the initial limitation of introspection and behavioral observations. Chapters 4 and 5 present the two main arguments that aim to avoid this problem and make a case for psychological altruism.

As I said in the introduction, egoistic explanations of human behaviors and motivations are common in philosophy and the social sciences. Some even take the egoistic view as "obvious" and "a matter of common sense"[20] (see Sober & Wilson, 1998, p. 2). But what are the reasons for believing in psychological egoism? It is not easy to answer this question, since not many authors have proposed an explicit defense of psychological egoism (see Berman, 2003). Nonetheless, the authors discussed in Chapters 4 and 5, who argue for psychological altruism, allow that psychological egoism has some initial advantages when compared to psychological altruism. For the sake of making clear the initial legitimacy of the debate on psychological altruism, I will present *four* points that can be used in support of psychological egoism. Before that, I should state clearly that *I do not endorse psychological* egoism: my goal is simply to help the reader to understand at least the initial plausibility of psychological egoism, which is required if we are to take the debate about psychological altruism as a legitimate and relevant debate.

The *first* point one can use to endorse psychological egoism is that it assumes the existence of a kind of motivation that is quite uncontroversial. People act in egoistic ways, and it is hard to make sense of their actions if we do not postulate egoistic ultimate desires. The idea that humans have ultimate egoistic desires is rarely, if ever, questioned by authors that assume that humans have beliefs and desires. So, psychological egoism does not assume a controversial sort of ultimate desire, while psychological altruism does.

The *second* point in favor of psychological egoism is that it, alone, offers explanations for all sorts of helping behaviors. Behaviors such as volunteering or donating an organ to a stranger,

[20] Blackburn (1998) adds that psychological "has a tremendous emotional power. If we believe it, we know the world, we are nobody's fool…; like conspiracy theorists, we have penetrated below the surface… we see the real face of human beings behind the mask" (p. 138). This psychological factor might also play a role in encouraging people to believe in this hypothesis.

which are usually regarded as altruistic, can be explained as means to egoistic ends. Altruistic behavior can be produced in many different ways (Feigin et al., 2014), and many are based on egoistic motives (Cialdini et al., 1981; Fehr & Fischbacher, 2004; Wynn et al., 2018). Egoistic motivation can explain even extreme cases, such as a soldier jumping in a grenade to save his comrades (Sober, 2013, p. 148): perhaps the soldier wanted to avoid the guilt he would feel otherwise, perhaps he wanted to be remembered as a hero, and so on (Batson, 1987, p. 66). For every scenario, there is always a possible egoistic explanation. So, if we accept the existence of egoistic ultimate desires, and they can explain all sorts of altruistic behaviors, then proponents of psychological altruism need to give us reasons for postulating yet another kind of ultimate desire.

The *third* point one can use to defend psychological egoism is that it can respond to contrary evidence from introspection. People may claim that they have introspective evidence that their motivation is altruistic, and that this should give *prima facie* support for psychological altruism. However, proponents of psychological egoism can respond to this by mentioning that ultimate desires can be unconscious. People may believe that they care for others ultimately, but that is just a believe they have, hiding the crude truth that they only help others for ultimate egoistic reasons. The idea that the egoistic character of our motivation is hidden from us is not a far-fetched claim: attributing false reasons for why we make certain choices is a common phenomenon well-known in psychology (see Johnsrude et al., 1999). If psychological egoism is true, it is plausible that mothers will believe that they have ultimate desires to benefit their children, as knowing the true egoistic source of their love would be distressing.

The *fourth* point in favor of psychological egoism is that, for any given helping behavior, the selection of egoistic motivation as its proximate mechanism is easier to explain than the selection of altruistic motivation. Individuals who help moved by egoistic motivation are more

likely to offer help only when helping is in their best interest, so they are less likely to be exploited when compared to altruistically motivated individuals. Individuals with altruistic motivation are more likely to provide maladaptive helping, where their own fitness is reduced. This is recognized in the evolutionary argument for psychological altruism, where authors only claim that altruistic motivation can be adaptive as a proximate mechanism for parental care. In the majority of helping behaviors, especially in cases of cooperation between non-related individuals, egoistic motivation seems to be the optimal proximate mechanism[21].

There might be all sorts of reasons for one to believe in psychological egoism. But the four points above give an overview of how one might justify a preference for psychological egoism over psychological altruism. The initial plausibility of psychological egoism is also endorsed by the authors discussed in Chapters 4 and 5. Daniel Batson (1991), known as one of the main proponents of psychological altruism, discusses evidence for psychological altruism provided by his colleagues and claims that they are not strong enough to undermine psychological egoism (p. 49). He states that "given that egoistic motives exist and altruistic motives may or may not exist, parsimony clearly favors an exclusively egoistic view" (Batson, 1991, p. 50; cf. Sober, 2013, p. 160). Sober and Wilson (1998) also recognize some advantages

[21] There is an important caveat about this fourth point. Firstly, we should distinguish two claims. One claim is that (1) for any particular helping behavior, egoistic motivation is more likely to be selected as its proximate mechanism. A different claim is that (2) all of our ultimate desires are egoistic. Even if (1) is true, it does not imply neither that (2) is true nor that it is more likely to be true. Egoistic desires require a complex representation of the self, and there are no reasons to assume that no desires can simply lack this complex representation of the self. It sems plausible that there are simpler desires, which are neither egoistic nor altruistic. Despite this caveat, however, this fourth point can still help psychological egoism, at least when directly compared to psychological altruism. One can ask: if humans solve all sorts of problems through egoistic motivation, why would they need a completely different sort of motivation when it comes to helping others? The fourth point is not a direct defense of the truth of psychological egoism, but it can, at least, be considered as a piece of evidence against psychological altruism.

of psychological egoism. Their argument aims to be strong enough to counter this initial advantage of psychological egoism — some commentators, such as Harman (2000), even complain that Sober and Wilson (1998) are too generous to psychological egoism. Wilson and Sober (2002) state that psychological egoism is *simpler* than psychological altruism (p. 725), and Sober (2013) goes through a set of common criticisms against psychological egoism, defending it from being dismissed too quickly.

The defense of psychological egoism presented here can be interpreted in two ways. In its *strong* form, this defense suggests that we should *prima facie* adopt psychological egoism as the most likely hypothesis. In this strong form, we leave the burden of proof to proponents of psychological altruism. However, we can also interpret the four points as suggesting a more moderate support for psychological egoism. In its *weak* form, the defense above simply states that psychological egoism is an initially plausible hypothesis. I claim that this defense grants at least the weak form. Importantly, though, we need to consider that the strong form seems to be assumed even by critics of psychological altruism. All the arguments for psychological altruism take the form of responding to psychological egoism as if it were prima facie the more plausible hypothesis.

In this chapter, I presented the many faces of altruism, showing a rich repertoire of technical accounts of altruism in the literature. The central topic was the standard account of altruistic motivation, which I presented in detail, highlighting the ways in which it differs from other accounts of altruism. Having a grasp of the different technical accounts of altruism in the literature will prove useful when I discuss the limitations of the standard account and how we have good reasons to abandon this particular way of thinking about altruism.

One of my goals in this chapter was to offer the reader a map of the different technical accounts of altruism, showing how they differ and how they can be useful for different purposes. In *Table 1*, I summarize all these accounts, showing their *necessary conditions*.

Table 1

The Necessary Conditions for the Different Accounts of Altruism

	Standard Account of Altruistic Motivation	Kitcher's Altruistic Motivation	Preference Altruism	Behavioral Altruism	Helping Altruism	Directed Altruism	Reciprocal Altruism	Evolutionary Altruism
Motivation to Help								
Ultimate desire to benefit others	•							
Conscious desire to benefit others		•	•					
Helping Behavior								
Helping behavior *benefits* others			•	•	•	•	•	•
Helping behavior is *costly* for the agent			•	•			•	•
Helping is *directed* to others in need		•				•		
Future Compensation								
Costs are *not expected* to be compensated		•		•				
Costs *are* compensated							•	
Costs are *not* compensated								•
Impacts on Fitness								
Costs are measured as effects on fitness							•	•
Benefits are measured as effects on fitness							•	•

The table above hopefully provides readers with an easy way to visualize the differences between the various accounts of altruism. In the next chapter, I delve deeper into the theoretical framework of altruistic motivation, exploring the different theories involved in the debate.

Chapter 3

The Altruistic Mind

3.1 The Framework of Altruistic Motivation

While the previous chapter distinguished the standard account of altruistic motivation from other kinds of altruism, this chapter discusses the philosophical framework underlying altruistic motivation, especially the concepts of mental state and desire. The standard account of altruistic motivation reduces altruism to certain mental states, namely, ultimate desires to increase the welfare of others. Underlying the notion of altruistic motivation there is a complex philosophical framework, with many assumptions, such as that beliefs and desires exist. It is crucial to have a clear understanding of these assumptions, making sense of the concepts and theories they involve. This is so especially considering the interdisciplinary character of the debate: if the arguments from authors from different areas are expected to be considered together, as part of the same coherent discussion, the conceptual framework of what I shall call the "altruistic mind" needs to be clearly stated.

The first element of the altruistic mind framework that I will discuss is the notion of mental states. In *Section 3.2,* I address the main theories in philosophy of mind, discussing how they influence our debate on psychological altruism. I first address Cartesian dualism, since much of the relatively recent discipline of philosophy of mind has been reactive to the dualist

answer to the mind-body problem[22]. After this, I address the radically anti-cartesian theory of logical behaviorism, which was very influential in the middle of the 20th century, and the identity theorists' reaction against the reductionism proposed by behaviorists. Finally, I address functionalism, which manages to combine the virtues of both behaviorism and identity theory. This brief historical and philosophical overview of the process leading to the contemporary dominant views on mental states highlights the background of the debate on psychological altruism. But it also shows the sort of theory on which the debate is based on.

In *Section 3.3*, I discuss the notion of folk psychology and the theory of eliminative materialism. Eliminativists claim that folk psychology, which contains terms such as *beliefs* and *desires*, is an outdated vocabulary that should be abandoned. Since the debate over psychological altruism is dependent on the assumption that beliefs and desires exist, eliminativism criticizes the core of psychological altruism. Traditional eliminativism poses a challenge for the debate, but it is too demanding, as it asks us to reject the whole vocabulary of folk psychology. A more sophisticated and pressing challenge is posed by a less global eliminativist view, which aims to eliminate specific terms. Should we eliminate altruistic motivation from our *scientific* vocabulary?

After the discussion about mental states, I move on to discussing the concept of *desire*. In *Section 3.4*, I present the main theories of desire in the literature. These theories dispute which features should be the essence of desire. The standard theory, usually adopted in the debate over psychological altruism, is that desires are motivations to act in ways to bring about what is desired. But alternative theories dispute this. The hedonistic theory claims that hedonistic states

[22] The mind-body problem is the problem of understanding what the relationship between our mental and physical states is. In other words, how the physical states of our nervous system interact with our mental states.

are what constitute the essence of desire. If this is true, then one can think about the opposition between altruism and egoism differently, as I will discuss in detail in Chapter 6. The reward-based theory of desire, in turn, claims that the essence of desire is to constitute a certain representation as a reward. I will not defend a particular theory of desire, but highlight how these different theories allow different approaches to the debate on psychological altruism.

This chapter goes from more general concepts within the altruistic mind framework to more specific ones. After discussing mental states and desires, in *Section 3.5* I discuss the notion of ultimate desire. How can we access ultimate desires? Is the contemporary approach to the debate, which makes ultimate desires obscure and inaccessible entities, warranted? These questions touch on important issues for the discussion presented in the following chapters.

Before moving on to the next section, I should mention that, since the scope of my discussion is restricted to altruistic motivation as an ultimate desire, this results in the omission of some important authors in the literature on altruistic motivation. One of them deserves special mention, namely, Thomas Nagel. In his influential *The Possibility of Altruism* (1970), Nagel articulates an account of altruism that is motivated by our *normative beliefs*. These normative beliefs, for Nagel, can motivate our actions independently of whether there is also a desire pushing us to do so. The view of beliefs motivating behavior is controversial, and I will not discuss it here[23].

[23] In this thesis, I adopt the standard view in philosophy, which says that desires have the primary role in motivation. This view is known as the Humean theory of motivation. As Hume (1739/1960) famously stated, "[r]eason is, and ought only to be the slave of the passions" (p. 415). For a discussion of anti-Humeanism and the idea of beliefs as the source of motivation, see Lewis (1988, 1996). Alex Gregory (2021) has recently argued that the distinction between desires and normative beliefs should be abandoned. He proposes an account of motivation where desires and normative beliefs are *identified*, claiming that they are simply "two different labels for a single thing" (Gregory, 2021, p. 2). This view has potential to shed some new light on the debate on psychological altruism.

A more central problem in Nagel's view of altruism, however, is his assumption of a *substantive* account of rationality. As Sober (2013) explains, Nagel assumes that one is justified to act when the act is universally acceptable and the ends are not morally objectionable (p. 156). Nagel's argument for the existence of altruistic motivation is based on the idea that "practical reasoning necessarily involves the adoption of an impersonal standpoint" (Darwall, 1974, p. 125). This account of reason diverges from the instrumental account of reason, which is standardly accepted today. A selfish person can do all sorts of terrible things and do so rationally (Sober, 2013, p. 156). Both proponents and critics of psychological altruism in the contemporary literature reject the substantial view of rationality. So, since the argument proposed by Nagel (1970) is situated in a very different tradition, relying on a different way of understanding the basic elements of the debate, I will not discuss his theory in this thesis.

3.2 From the Soul to Neurons: Mental States in Philosophy of Mind

One of the main questions in the history of philosophy concern the nature of our mental states. What is our mind made of? Where is it located? How does it interact with our body? How can we know the mental states of other people? In the first half of the seventeenth century, René Descartes (1596–1650) defended the view that our mental states are radically different from our physical states at a metaphysical level. Descartes (1641/1985) argued that mind and body are properties or states of different *substances* (p. 50). The mental substance (*res cogitans*) and the physical substance (*res extensa*) do not interact in the way that parts of the same mechanism interact. Our mental states have a completely different way of existing and cannot be *reduced* to

physical states. This view is known as *mind-body dualism*. In this view, mental states are simply not part of the material world.

Although the idea of the mind as a non-physical substance was present since the origins of philosophy, being a central idea in Plato and the Christian tradition, it is in Descartes that it finds its most prominent formulation (Ludwig, 2003, p. 13). Descartes's way of thinking about the mind, relying on introspection and giving the mind a primacy over the body, has dominated philosophy of mind for the majority of its modern history (Lycan, 2003, p. 47).

If mental states are not part of the material world, as Cartesian dualism claims, then the mind is not an object we can investigate empirically. But as empirical sciences became more and more successful in explaining nature, naturalistic explanations of the mind became more appealing. In the second quarter of the 20th century, the philosophical tradition known as Logical Positivism offered the philosophical background for a scientifically oriented account of mental states. Cartesian dualism gradually appeared as a metaphysical extravagance that should be substituted by another theory more at home with scientifically oriented naturalistic views[24].

In contrast with Descartes, who relied on a highly introspective method, logical positivists strongly rejected any knowledge that was not based on empirical observation or logical inferences. In their view, for any sentence to be meaningful (to have cognitive meaning), it needs to be either verifiable or analytical (see Ayer, 1971). In an early paper of the logical positivist movement, Moritz Schlick (1930/1959) says that "[w]herever there is a meaningful problem one can in theory always give the path that leads to its solution", and this path to the

[24] Although the Cartesian view on the nature of mental states is not popular among contemporary philosophers, it should be noted that some authors in the contemporary philosophy of mind have argued that some aspects of our mental lives, which constitute our subjectivity, are unexplainable by a purely objective inquiry (Nagel, 1974; Jackson, 1982).

solution always requires the "occurrence of a definite fact that is confirmed by observation" (p. 54). Empirically meaningful sentences should be directly based on observation or derived deductively from other sentences based on observation.

In light of Logical Positivism, Cartesian dualism was clearly flawed: it does not offer any means of verification and it is not analytical. For logical positivists, the mind is a phenomenon like any other natural phenomenon. So, if we are to say anything meaningful about the mind, our claims should be verifiable, at least in principle[25]. Thus, for the logical positivists, the dualist view, which conceives mental states as non-physical entities, should be rejected not for being false, but worse — for being *meaningless* (see Burge, 1992).

Logical Positivism paved the way for new naturalistic approaches to the mind. The account of mental states directly inspired by logical positivists was *logical behaviorism* (Graham, 2019). Logical behaviorists proposed that the right approach to understanding the mind was to focus only on *observable* states. For them, any meaningful statement about our mental states can be translated into statements about our *behavioral dispositions* without loss of meaning (Fodor, 1981, p. 3). In this view, to desire x is, for example, to be disposed to present a set of behaviors towards the realization of x, such as acting in ways that bring about x, the production of utterances regarding preferring x, and so on. For behaviorist philosophers, in order to be meaningful, beliefs and desires "can't refer to unobservable events taking place inside a person", and "the meaning of sentences invoking these terms must be analyzed in terms of conditional sentences specifying how someone would behave under various circumstances" (Stich & Nichols, 2003, p. 236).

[25] The bold account accepting only empirically verifiable and analytical sentences as meaningful was characteristic of the early logical positivism, being abandoned in the late stage of the movement (e.g., Hempel, 1950/1959).

Ryle's (1949/2009) *The Concept of Mind* is arguably the most important philosophical work from the behaviorist tradition. In this book, Ryle offers a diagnosis of what is wrong with the Cartesian view. He explains that such an account follows from a particular mistake, which he calls a "category mistake". This mistake occurs when we apply concepts to categories they do not belong to. In a famous example illustrating what a category mistake is, Ryle asks us to imagine a foreigner visiting Oxford or Cambridge. On his first visit to the new university, after seeing "a number of colleges, libraries, playing fields, museums, scientific departments and administrative offices", the foreigner asks: "'[b]ut where is the University?'" (Ryle, 1949/2009, p. 6). The foreigner used "university" as a concept that belongs to the same category as "colleges", "libraries", and "playing fields", thus committing a category mistake.

Analogously to the foreigner who fails to use the concept of university correctly, Ryle claims that dualists represent our mental life as if it were of the same logical type as our physical life. Mental states, for them, exist in some way as physical objects exist. However, since Descartes rejected the reduction of the mental to the physical, dualists have to postulate another substance where mental states can exist like our physical states exist in the material world. Ryle (1949/2009) calls Descartes's view of the mind "the dogma of the ghost in the machine" (p. 5). Rejecting the Cartesian view, Ryle argues that mental states are merely ways we use to talk about behavioral dispositions.

What would a behaviorist say about the contemporary debate on psychological altruism? Probably not much. That is because the contemporary debate postulates the existence of mental states that do not have any necessary behavioral effect. Ultimate desires to increase others' welfare might be unconscious and might not have any effective motivational power to cause helping actions. So, from a behaviorist perspective, the debate on psychological altruism is

grounded on a mistaken way of thinking about mental states. Behaviorists would possibly propose a very simple solution to the debate on altruistic motivation: we should get rid of it in the same way that we did with dualism and other metaphysical doctrines.

Although the behaviorist criticism of the dualist view was relatively successful, logical behaviorism[26] also had its own flaws. By denying the existence of mental contents, the behaviorist approach departed from the most basic common sense. Furthermore, the method of reducing mental states to observable behavior was very limited. Behaviorists' "meaning analyses typically turned out to be either obviously mistaken or circular — invoking one mental term in the analysis of another" (Stich & Nichols, 2003, p. 236). Neglecting mental states in their explanation was a problem, since "[m]ental causes typically have their overt effects in virtue of their interactions with one another, and behaviorism provides no satisfactory analysis of statements that articulate such interactions" (Fodor, 1981, p. 5).

But while behaviorism has several flaws, its rejection should not entail the rejection of a naturalistic approach to the mind — doing so would be throwing the baby out with the bathwater. Philosophers needed an alternative naturalistic approach to mental states that allowed us to talk about mental states without having to return to the obscure language of dualism. Such an alternative was offered by Place (1956), Feigl (1958), and Smart (1959), who developed a

[26] It is important to distinguish the philosophical movement of logical behaviorism from behaviorism in psychology. The term behaviorism was coined by the psychologist John Watson (Schneider & Morris, 1987). Originally, this term represented a methodology, which recommended what sort of phenomena should psychologists focus on. This normative theory, nowadays called methodological behaviorism, claims that, since mental states are private, unobservable the analysis of mental states relying on introspection is not a sound methodological strategy (Graham, 2019). A different use for the term "behaviorism" is present in the research program known as *psychological behaviorism*. This program aims to explain behaviors in terms of stimuli and responses. B. F. Skinner is the most influential author in psychological behaviorism, and his theory became known as *radical behaviorism* (Schneider & Morris, 1987).

new theory of mind that could overcome the challenges raised against behaviorism. This new physicalist theory of mind became known as “mind/brain theory”, “reductive physicalism”, “type physicalism”, or simply “identity theory”.

For identity theorists, accepting the existence of inner mental states does not entail the acceptance of dualism (Place, 1956). One can be a physicalist and still accept the existence of mental states — as long as there is nothing in these mental states that go beyond physical states. As Place (1956, p. 48) explains, when someone sees lightning, we can say that what she observed was a motion of electric charges. Analogously, when she reports a given mental state, what she is reporting is just some physical state in her brain. Identity theory *identifies* mental states with physical states of our nervous system. The reduction of our mental states to our neurobiological states offered a way of integrating our understanding of the mind into the scientific framework. Identity theory provided a simple way to account for the correlations between mental states and brain states, such as those that we observe in cases where the brain is damaged. The development of cognitive sciences, with the emergence of modern neuroscience, contributed to the acceptance of this theory.

Different from behaviorism, the identity theory framework allows us to make sense of the debate on psychological altruism as it is debated in the contemporary literature. Altruistic desires can exist even if they are unconscious and do not produce behavior, as long as they are there, physically, in the brain. But how can we access these desires? In Section 2.3, I will discuss how some authors articulate a physicalist theory of desire.

Identity theory became a prominent view in the philosophy of mind. But not for long. Although this theory allowed us to overcome challenges raised against behaviorism, it also had some flaws. Perhaps the main problem with identity theory is that it established an overly

exclusive account of mental states. Putnam (1967) proposed an influential argument against the mind-brain identity theory by pointing out the possibility that two individuals might share the same mental state without sharing the same physical state. That is, identity theory does not consider that mental states might be *realized* by multiple physical states. Imagine, for example, an intelligent creature from another planet, which behaves and uses language just like us, but has a completely different physical constitution. A human and this creature might experience the *same* mental states, such as an ultimate desire to increase others' welfare. However, since this creature is constituted differently from humans, this altruistic desire will be realized by different physical states[27]. If identity theorists identify such a desire with certain states of our nervous system, they would have to deny this possibility. The mental states of the human and the creature are not the same, for they do not have the same physical states.

Logical behaviorism and identity theory are important theories in the history of philosophy of mind. They provide important intuitions about the nature of mental states that cannot be ignored. However, they have significant limitations. After the criticism directed to both identity theory and logical behaviorism, a new theory called *functionalism* was proposed as an alternative. The basic idea of functionalism is that mental states should not be defined in terms of what they are *made of* (as, for example, the definition of water as H2O), but instead based on the *function they have*. Functionalists accepted the identity theorists' claim that mental states are physical states. However, not falling into the same overly restrictive definition of

[27] Perhaps some readers would find a more realistic example more appealing. Consider a human and an octopus. Arguably, both are capable of experiencing pain. However, the realization of pain in these two organisms will follow from different physical states. Furthermore, if we consider the phenomenon of neuroplasticity, where, for example, the function from a damaged part of the brain is realized by another part, it seems reasonable to presume that mental states in different individuals of our species might be realized by different physical states.

identity theorists, functionalists argue that mental states are not *defined* in terms of their neurobiological features. Instead, mental states are defined in terms of their causal relations — our mental states are occupants of certain *causal roles* (Lewis, 1972).

Functionalists refined the original intuitions of identity theorists, avoiding an overly restrictive formulation. But they also refined the original intuitions of behaviorism. Functionalists accepted the logical behaviorists' claims that our mental states are determined by their causal relations to our sensory stimulation and its behavioral outputs. However, not committing the same mistake as behaviorists, functionalists accept that our mental states are *also* determined by their causal relations to other mental states. Functionalism combined the positive aspects of these two theories, avoiding their problems (Fodor, 1981, p. 10).

Both identity theory and logical behaviorism are *reductive* theories. They reduce mental states to physical states and to behavioral dispositions, respectively. Functionalism, however, is not necessarily reductive. There are reductive and nonreductive functionalistic accounts. *Realizer functionalists* (e.g., Lewis, 1972) identify mental states with *whatever occupies the causal role* attributed to that mental state. They reduce mental states to the physical states realizing the function — e.g., in humans, pain might be C-fibers stimulation and the neurobiological processes associated with it. By contrast, *role functionalists* (e.g., Putnam, 1975) identify mental states with *the role itself*, not reducing mental states to certain physical states. In this account, to be in pain is a property of having a certain functional state, independently of how this function is realized. Both accounts consider that mental states are defined based on the functional role they have, but they diverge on what mental states should be identified with (see Bennett, 2007).

Functionalism became the most popular theory of mind in contemporary philosophy, and the debate on psychological altruism is situated in the tradition of functionalism. It is easy to see how the debate on psychological altruism fits well in a functionalism framework. The very idea of "instrumental" and "ultimate" desires assumes that mental states can causally interact with each other. Beliefs and desires, as propositional attitudes, have causal relations both to behaviors and other mental states. They interact in inferential processes in our decision-making, which are analogous to the inferences in a syllogism (Fodor, 1981, p. 183). Our actions in the world are based on the results of this inferential process. However, this functionalist view of the interactions of our beliefs and desires only follows if we assume that these mental states *exist*. But are we warranted in making such an assumption? I discuss this in the next section.

3.3 Folk Psychology and Eliminativism

The previous section discussed different views about mental states. But although these different views dispute the nature of mental states, they share the set of terms of what is known as "folk psychology". The different theories of mind allow us to use terms such as "pain,", "fear", and "desire" to describe our mental states. Eliminative materialism, however, claims that this vocabulary we use to refer to mental states is flawed and should be abandoned. In this section, I discuss the problems of folk psychology and the response given by eliminativists. I argue that, although traditional eliminativism is too radical, scientific eliminativism, which excludes particular terms from scientific vocabulary, offers a reasonable response to problematic folk-psychological terms.

In the middle of the 20th century, Sellars (1956) proposed that our folk theory of the mind is not based on direct access to how our minds work, but on a theoretical framework. The psychological platitudes accepted by common sense, such as "pain tends to cause one to avoid the painful stimulus", are part of a theory. This idea, known as "theory-theory", states that folk psychology is a theory we have to explain and predict behaviors and mental states. On this account, folk psychology is a term-inducing theory, giving "ordinary mental state terms their meaning" (Stich & Nichols, 2003, p. 237). In this view, beliefs and desires are part of a theory, and mindreading is a form of theoretical reasoning (see Hutto & Ravenscroft, 2021).

The assumption that folk psychology is a theory allows us to make some interesting comparisons with other folk theories. Consider, for example, the theory we use daily to explain and predict the phenomena in the physical world, which we may call "folk physics" (see Dennett, 1989). If we adopt the theory-theory view, folk psychology is to the human mind what folk physics is to the physical world. They both are theories we use to predict and explain certain phenomena. This comparison highlights that, if we can say that our folk physics is wrong, we may be able to do the same to folk psychology. In the same way that many of our intuitions about the physical world are revealed to be false by physicists, the widely accepted intentional vocabulary we use daily might also be simply wrong (Churchland, 1981, p. 72). Folk psychology is *falsifiable*. Some authors argue that, in fact, folk psychology is a bad theory and should be rejected. These authors are known as *eliminative materialists*.

Traditional eliminative materialism, whose most prominent proponents are Paul Churchland (1981, 1985) and Patricia Churchland (1989), claims that common sense's understanding of what is going on in our minds is profoundly mistaken. They argue that, on close examination, folk psychology describes our mental states very poorly, particularly when we take

into account the neuroscientific knowledge available today. More than getting things wrong, folk psychology also leaves a large range of mental phenomena unexplained: "questions concerning memory and learning, motivation, dreams, coma, the dementias, pain disorders such as Congenital Insensitivity to Pain are entirely overlooked by [folk psychology]" (van Rysewyk, 2016, p. 75).

The fact that folk psychology presents a set of explanatory deficiencies does not surprise eliminativists. After all, they say, folk psychology has remained basically unchanged for thousands of years (Churchland, 1981, p. 73). Many other folk theories have been eliminated in the meantime, such as theories of matter, medicine, cosmology, etc. Thus, it seems plausible to expect that the same would happen to folk psychology. Eliminativists claim that there is nothing special about our current mentalist vocabulary. In the same way that we rejected Ptolemy's geocentric model of the universe or the phlogiston theory of combustion, we should also reject folk psychology. "[T]he principled displacement of folk psychology is not only richly possible, it represents one of the most intriguing theoretical displacements we can currently imagine" (Churchland, 1981, p. 90). Perhaps the only reason why we do not see that folk psychology should be eliminated is the fact that its object, the human mind, is so complex that we still do not have sufficient knowledge to appreciate why this theory is wrong.

Eliminative materialism has a dramatic effect on the debate on psychological altruism. As Wilson and Sober (2002) comment, if the "Churchland-style eliminativism" is right, then the whole debate on psychological altruism is a non-existent problem (p. 724). If folk psychology turns out to be wrong, so that we should abandon it, then we should do the same with the whole framework of the debate on psychological altruism. This debate is entirely based on folk psychology in its way of understanding altruistic and egoistic motivations. It is accepted that we

have beliefs and desires, which is a folk-psychological view. Therefore, the debate on whether eliminativism is right has direct implications for the debate over psychological altruism. As Sober and Wilson (1998) state, "[i]f science someday establishes that beliefs and desires do not exist, …then we expect the debate about psychological egoism and altruism to be tossed on the rubbish heap of history" (p. 208).

The idea of *completely* eliminating folk psychology is very radical and is usually received with skepticism. Talking about beliefs, desires, pains, and so on, is extremely useful. This is true not only in everyday life but also in the social sciences. Psychology, anthropology, sociology, criminology, and many other areas rely on the assumption that humans have beliefs and desires. Although discoveries in neuroscience have an impact on the way we use intentional vocabulary, there is no evidence that these areas would do better if they substitute their vocabulary for neuroscientific terminology.

Traditional eliminativism seems to be too harsh a response to the flaws of folk psychology. But we can consider more moderate forms of eliminativism. We can consider, for example, the possibility of eliminating particular concepts. For example, there is much discussion about the elimination of the term "pain" (Dennett, 1978; Hardcastle, 1997). One can argue that this term should be eliminated, for being inconsistent, but accept, at the same time, that other terms from folk psychology should be maintained. Such eliminativism, directed to particular terms, seems more acceptable than traditional eliminativism. But I want to address another form of eliminativism, which is yet more moderate, namely, *scientific eliminativism*.

Instead of eliminating folk psychology as a whole, or particular terms from it, we can consider eliminating particular terms from a specific discourse. Jennifer Corns (2016), for example, argues that, while the traditional eliminativism for the term "pain" is not plausible,

scientific eliminativism for this term is reasonable. We can be in accordance with keeping the term pain in everyday discourse and defend its elimination from scientific discourse.

Traditional eliminativism, if accepted, would undermine the whole debate on psychological altruism. However, this is a theory that does not find much support today. In moderate versions, however, such as scientific eliminativism, it becomes a more serious threat to the debate. Should we eliminate the notion of "altruistic motivation" from the scientific vocabulary? Although I will not argue for this, I consider this to be a plausible alternative to the problems of the standard account of altruistic motivation. Once we consider the normative elements of the ordinary use of "altruism", for example, there might be good reasons for one to claim that we should eliminate this term from scientific research.

So far, I have discussed different theories of mind and some implications for the way we think about altruistic motivation. The standard account of altruistic motivation, as I have shown, is dependent on a series of theoretical assumptions that are still debated. Now, I shift the discussion from mental states to *desires*, which is the most important mental state in our discussion on altruistic motivation.

In Chapter 2, I explained that desires and beliefs are propositional attitudes with different directions of fit. They share the same kind of content and, thus, interact in decision-making. Desires are understood to have a motivational component, which guides our actions in the world. In the debate on psychological altruism, it is also assumed that altruistic and egoistic desires motivate individuals to act in certain ways (Sober & Wilson, 1998, p. 208). But there are different ways in which desires might fulfill this motivational role (Schroeder, 2020). Different theories explain differently how this motivation occurs. In the next section, I will present the main theories of desire in the literature, discussing some of the implications for our debate.

3.4 Theories of Desire

I start by presenting the *action-based theory* of desire, which is the standard theory of desire in the literature. This theory claims that the fundamental feature of desire is its motivational component. As Smith (1994) explains, in this theory desiring p is to have certain dispositions to produce, under certain conditions, actions that would bring about what is desired. The action-based theory is a functionalist theory. It defines desires in terms of the causal role they occupy in decision-making (Smith, 1994, p. 113). Although desires may have different roles, if we accept the action-based theory "there is just one role definitive of desiring, and that is of engaging the mental machinery in such a way as to tend to bring it about that P" (Schroeder, 2004, p. 11). The action-based theory is widely accepted by philosophers and can be considered the standard account of desire in the literature (see Schroeder, 2004, p. 10; see also Goldman, 1970, p. 112).

The action-based theory can be formulated in different ways. A simple definition could state that "[t]o desire that P is to be disposed to bring it about that P" (Schroeder, 2004, p. 11). However, this is very inclusive and vague. We have plenty of dispositions that make us behave in certain ways that bring about states of affairs we do not desire. I am disposed to blink if someone claps in front of my face, but I do not have a desire to do so. A better definition, which could avoid the overly inclusive nature of the previous one, states that "[t]o desire that P is to have a *mental representation* [emphasis added] that P which plays a certain causal role, namely, that of disposing one to bring it about that P" (Schroeder, 2004, p. 24). In this definition, desires are not just any disposition, but a *representational* state.

The idea that desires motivate us to act in certain ways is certainly consistent with the commonsense view of desire: in common parlance, we usually agree that we are likely to behave in ways that tend to bring about the things we desire. But it is not obvious that the motivational component should be the *essential* feature of desire. Some authors have rejected the action-based theory (Schroeder, 2004; Arpaly & Schroeder, 2014; Schroeder, 2020). I will address one of these objections, which was raised by Strawson (2010).

In a thought experiment, Strawson (2010) postulates creatures that he calls "weather watchers" (p. 251). As their name suggests, these creatures observe the weather. But these creatures are, by their nature, incapable of producing any kind of behavior. Notwithstanding this limitation, they still *desire* certain phenomena to occur, such as sunny days. The weather watchers have a rich subjective life, with complex emotions, although they cannot act[28]. Both Strawson (2010) and Schroeder (2004) believe that the fact that we can imagine these creatures as holding genuine desires highlights that making dispositions to act the essence of desire is a mistake. If these creatures are capable of desiring, then maybe the motivational component is not the essential feature of desire.

Consider now a different theory of desire. This theory can easily account for the desires of the weather watchers. This is the *hedonistic* theory of desire, which emphasizes *feelings* rather than dispositions to act (see Strawson, 2010, p. 266). The hedonistic theory of desire can be

[28] Another example could be paralyzed persons. They desire, although they cannot act. Strawson's (2010) thought experiment aims to avoid complications regarding what a paralyzed person can do. For example, one can say that people in this condition *try* to act, and that trying to act is already an act (Strawson, 2010, p. 271). But I believe that a similar problem can be raised against the weather watchers. The psychology of weather watchers is either similar or different to ours. If their psychology is *similar* to ours, we will imagine them trying to act when they desire, just like paralyzed persons. But if their psychology is unlike ours, then we cannot know what it is like to be a weather watcher, and we have no reasons to believe that they have desires.

defined as stating that "[t]o desire that P is to be so disposed that one will tend to feel pleasure if it seems that P, and/or displeasure if it seems that not-P" (Schroeder, 2004, p. 27). If we adopt this theory, then we can say that, although the weather watchers do not have any dispositions to act, they still have dispositions to *feel*. Thus, they can have desires.

Just like motivation, the tendency to feel pleasure is one of the main features associated with desire. In common sense, it is usually accepted that if a desire is realized, a pleasurable sensation would be expected. "Tendencies to pleasure and displeasure are taken to be very powerful evidence of desire, often overriding other sources of evidence such as statements of what one desires…, behavioral evidence, and so on" (Schroeder, 2004, pp. 28-29). The question, however, is whether we should make tendencies to feel pleasure and displeasure the *essential* features of desire[29].

Schroeder (2004) raises an objection to the hedonistic theory. He claims that it does not offer a proper explanation of the motivational component of desire, for the neurological basis of pleasure and displeasure seems to have a small contribution to the brain areas responsible for producing movement (Schroeder, 2004, p. 127). Schroeder (2004) claims that neuroscientific evidence suggests a different role for pleasure and displeasure: "[p]leasure and displeasure are best interpreted as sensory representations of the difference between actual and expected desire satisfaction, and are thus a type of sense experience whose object is the subject's own desires" (p. 31). In Schroeder's view, hedonic states represent changes in desire satisfaction analogously to how vision represents the environment in front of us. So, considering that "pleasure and

[29] The hedonistic theory of desire offers an interesting new perspective on the debate on psychological altruism. It allows us to consider pleasures and pain to be the ultimate cause of motivation without having to accept psychological egoism. I will discuss this in Chapter 6.

displeasure represent desires, they cannot be even partially constitutive of desire" (Arpaly & Schroeder, 2014, p. 124).

After criticizing the action-based and the hedonistic theories of desire, Schroeder (2004) proposes a new theory. In his *Three Faces of Desire* (2004), Schroeder says that the *three* main components associated with desire are motivation, hedonistic states, and reward/punishment. The theories addressed previously make either motivations or hedonistic states the essence of desire. But following Dretske (1991), Schroeder gives primacy to the third face of desire. He proposes the *reward-based* theory of desire, which makes rewards the essence of desire. In this theory, "to have an intrinsic desire regarding it being the case that P is to constitute P as a reward or a punishment" (Arpaly & Schroeder, 2014, p. 127). More precisely,

> To have an intrinsic (positive) desire that P is to use the capacity to perceptually or cognitively represent that P to constitute P as a reward. To be averse to it being the case that P is to use the capacity to perceptually or cognitively represent that P to constitute P as a punishment (Schroeder, 2004, p. 131)

Schroeder (2004) begins his defense of the reward-based theory by stating that "there is a link between desire and reward: desires determine what counts as a reward and what counts as a punishment for an organism" (p. 15, 67). But even when we recognize the role of rewards in desiring, making reward the *essence* of desire seems odd. Dispositions to act and hedonistic states seem to be better candidates to be the essence of desire. However, despite the commonsense view, Schroeder argues that rewards and punishments should be taken as the fundamental features of desire because both motivation and hedonic states are *consequences* of the reward system. Following Morillo (1990), who tries to articulate an understanding of desire using neuroscientific evidence, Schroeder adopts an empirically oriented approach to support his

reward-based theory. The strength of his case for the reward-based theory stems from its neuroscientific support, which I discuss in the next section.

Here is a summary of Schroeder's reward-system theory. Schroeder (2004) starts by claiming that the motivational aspect of desires should be understood as a *consequence* of the reward system. Research shows that two dopamine-releasing structures (ventral tegmental area, or VTA, and substantia nigra pars compacta, or SNpc) constitute the neural basis of our reward system (Schroeder, 2004, pp. 36, 50). These areas are responsible for a certain kind of learning (contingency-based learning) and can modify behavioral dispositions, which makes them central to the production of voluntary movement[30] (2004, pp. 54, 115). Parkinson's disease affects dopamine-releasing cells in the SNpc and has well-known effects on one's movement production, going from tremors to the total absence of movement. This disease illustrates the link between the reward system and movement production. Moreover, the basal ganglia, which is responsible for the selection of action, is also strongly influenced by the neural basis of reward. Considered together, the points mentioned here make the reward system a plausible neurological basis for desiring, accounting for how it can lead to motivation.

The second step in Schroeder's argument is to show that hedonistic states are also consequences of the reward system. Research suggests that the perigenual anterior cingulate cortex (PGAC) is responsible for the capacity to feel pleasure and displeasure (Schroeder, 2004, pp. 36, 77). PGAC is completely distinct from the reward system structures (VTA and SNpc) —

[30] The process that makes a proposition to be constituted as a reward is a kind of learning. Neuroscientific evidence shows that the main function of the reward system is producing a special kind of learning, called contingency-based learning (Schroeder, 2004, p. 61). "For an event to be a reward for an organism is for representations of that event to tend to contribute to the production of a reinforcement signal in the organism" (Schroeder, 2004, p. 66). For this reason, the reward-based theory is also called "learning-based theory" (Schroeder, 2020).

the neural basis of reward is not identical to the neural basis of pleasure. This dissociation is reflected in the fact that rewards can occur without pleasure, as in some cases of drug addiction (Schroeder, 2004, p. 60). Furthermore, pleasure can also occur independently of rewards (Schroeder, 2004, p. 81). As I mentioned before, Schroeder (2004) argues that pleasure and displeasure are better seen as sensory experiences indicating whether our desires are satisfied or not (p. 89). This picture makes hedonistic states no longer a good candidate for being the essence of desire. Both the motivational and the hedonistic aspects of desire can be seen as consequences of or secondary to the reward system.

Schroeder (2004) offers a theory where desires are understood as a specific neurobiological phenomenon. In his view, desires are "a meaningful, unified, scientific entity" (Schroeder, 2004, p. 6). This view is shared by Schulz (2018), who sees beliefs and desires as part of our cognitive architecture, not merely "a point about the kinds of explanations humans engage in as part of everyday life to make sense of each other's behaviors" (Schulz, 2018, p. 29). Sterelny (2003) calls this the Simple Coordination Thesis: the reason why our folk psychology is so effective is that it is *true* — it actually describes facts about the "wiring-and-connections" of our mind.

Schroeder (2004) and Schulz (2018) resist views of desires as merely useful abstractions. They oppose views such as Dennett's intentional stance. To Dennett (1989), we can adopt an interpretative strategy to *explain* and *predict* the behavior of a "system" (which might be a human, an animal, a machine, etc.) by attributing beliefs and desires to this system (p. 17). In this account, to believe in p is simply to be disposed to behave in rational ways given the truth of p and given the rest of the other beliefs and desires that this individual has (Dennett, 1989, p. 50). This view does not demand representational mental states. Beliefs and desires, in this account,

are *not* descriptions of the mechanisms in one's mind. Even inanimate objects can be said to believe and to desire: a thermostat, for example, "will turn off the boiler as soon as it comes to believe the room has reached the desired temperature" (Dennett, 1989, p. 22). The success of our use of beliefs and desires, in this account, is no metaphysical evidence of their existence in our minds — it is only evidence that this is an efficient way of representing agents (Lycan, 2012, p. 201).

If Schroeder (2004) and Schulz (2018) are right, desires are natural kinds with a neurobiological basis, not merely useful abstractions. This approach brings the debate on psychological altruism closer to the empirical sciences. Taking altruistic motivation to be a natural kind allows us to make sense of altruistic motivation as a trait subjected to natural selection, which will be discussed in Chapter 5.

The three theories discussed in this section assume that there is a single essence of desire, diverging only about which feature is essential. But we should also consider that it is possible to reject the assumption that there is a single essence of desire. One could follow Strawson (2010), for example, and consider *two* different features that are sufficient for desires (p. 283). This "disjunctive" theory of desire could state that a desire for p is to have *either* a disposition to act in ways that bring about p *or* to have a disposition to feel pleasure when p seems to be the case. But we can go further and exclude sufficient features from our theory of desire.

Holistic theories of desire (e.g., Lewis, 1972) reject the need for both necessary and sufficient features for desires (see Schroeder, 2020). In this approach, we have a *cluster* of properties associated with desires, but none of them are necessary or sufficient to realize a case of desire. The set encompassing all desires might be composed of states that share only a *family resemblance* (Wittgenstein, 1953/1968, p. 32). In this approach, we have a *cluster* of properties

associated with desires, but none of them are necessary or sufficient to realize a case of desire[31]. Holistic theories seem to fit nicely with the common-sense view, which applies the term "desire" to a set of different states. It also avoids all the problems raised against the other theories. This theory, however, in inherently vague, making its use very challenging.

3.5 Ultimate Desires

The three initial sections of this chapter discussed theories of mental states. In the previous section, I narrowed down the focus to desires, discussing the different theories of desire in the literature. Now, in this last section, I narrow the focus further, discussing the subset of desires that constitute altruistic motivation, namely, *ultimate desires*. The challenging nature of the debate on psychological altruism follows precisely from the difficulties in accessing ultimate desires. Introspection and the observation of behaviors are not epistemically reliable means of accessing ultimate desires. In this section, I discuss the problem of accessing ultimate desires.

Before asking how we can access ultimate desires, we should ask how we access desires in general. The philosopher Alvin Goldman, (1970) claims that "[t]he inference from observable facts concerning human beings to the proposition that they have wants [desires] is not a deductive inference, nor an inference by enumerative induction…; rather it is an inference to the best explanation" (p. 114). The claim that we or others have a given desire is an inference based on the evidence we have, including introspection, observation of behaviors, and any other source of evidence. There are a set of features that often accompany desires, regardless of the theory of desire we adopt. A propensity to act in ways that lead to the realization of x might indicate a

[31] I explain the notion of family resemblance in more detail in chapter 8.

desire for x; a propensity to feel pleasure when it seems that x might also indicate a desire for x; and so on. Based on the observation of these features, we can, sometimes, infer that a certain desire exists.

Remember that the deadlock problem in the debate on psychological altruism follows from the fact that we cannot access whether a given desire is ultimate. But can we use inferences to the best explanation as a means to know whether a desire is instrumental or ultimate? To some degree, we can. There are at least two distinct mechanisms forming instrumental desires. One is a rational deliberate inference and the other is an unconscious process of inference. We can apply different methods for each of these two mechanisms to infer whether a desire is instrumental.

Consider the first mechanism. In this mechanism, the instrumental desire is formed by our conscious calculation. For this mechanism, we have introspective access to the desire formation, so we can know whether it is instrumental. For example, if I desire a cup of coffee and think that I can get one if I walk to the cafeteria, I can form an instrumental desire to go to the cafeteria. In this case, I will have access to the status of my desire to go to the cafeteria as instrumental, since I was conscious of the process producing this instrumental desire.

Consider now a second mechanism forming instrumental desires. In this mechanism, instrumental desires are formed by an unconscious inferential process. We do not know all of our desires, and the inferential mechanism linking one desire to the other, relying on our beliefs, can occur without our deliberate intention. In these cases, is it possible to have good reasons to believe that a given desire is instrumental? In some cases, yes. Imagine, for example, that I have a desire for x and x is a means to y. Imagine that I desire y ultimately, and I only desire x instrumentally. Now, consider that this desire for y is unconscious — all I know is that I desire x. In this case, how can I learn that my desire for x is instrumental? A scenario that could allow me

to make the inference that my desire for x is instrumental is one where the desire for y is satisfied and I no longer hold the (unconscious) desire for y. When I stop desiring y, my desire for x will cease. When this happens, I can *infer* that my desire for x was instrumental to an unconscious desire for y (see Batson, 1991, p. 65).

The two methods above are limited. In both methods, we can only identify when a particular desire is *instrumental*. The desire to go to the cafeteria is instrumental, but what about the desire for a cup of coffee? It might also be instrumental to further unconscious desires. I might have good reasons to believe that my desire for x is instrumental for y, but I would still not know whether y is instrumental to a further desire. The methods above can show us when some desires are *instrumental*, which is important and can advance our understanding of our motivations, but they cannot tell when a desire is *ultimate*.

The challenge in the debate about psychological altruism is mainly concerned with cases in which our inferential process is unconscious. So, the second method used to identify instrumental desires above is of particular interest. But this second method has further limitations we should consider. We can know that a given desire for x is instrumental only when another desire for y (to which x is instrumental) *ceases to exist*. This can occur when we, for example, satisfy a desire for a particular object and no longer desire it. However, in the debate on psychological altruism, we have to consider deep egoistic desires, which are unlikely to cease. For example, the desire to increase one's pleasure and reduce one's pain. These desires are very stable, and it is not clear how one could produce a scenario in which these would no longer be present. Thus, we cannot compare a scenario in which we have these egoistic desires with a

scenario in which we do not have them, making it very difficult for us to apply the abductive method in the debate on psychological altruism[32].

Let me now address a different issue regarding ultimate desires. This issue concerns two different ways of conceiving the conditions for accepting a given desire as ultimate. There are two radically different views in the literature about how we access ultimate desires. I will call them the *hard-access view* and the *easy-access view*. The rest of this section aims to present these two views and explain their consequences for the debate on psychological altruism.

Consider first the *hard-access view*. In this view, we accept that ultimate desires are few and very hard to access. It assumes that the vast majority of our desires are instrumental. This view offers a *prima facie* advantage for psychological egoism, since the set of ultimate desires assumed by psychological egoism is restricted and uncontroversial (most accept that the desire to avoid pain is not instrumental, for example). Apart from the standardly accepted set of ultimate desires, the hard-access view will require robust evidence justifying the claim that other desires are ultimate. The hard-access view leaves the burden of proof for proponents of psychological altruism, who want to defend the existence of a new, controversial ultimate desire. This seems to be the standard view adopted by proponents and critics of psychological altruism, and it explains the structure of the debate, in which proponents of psychological altruism are the ones expected to defend their view against psychological egoism.

By contrast, we can also find an *easy-access view* of our access to ultimate desires. This second approach assumes a very permissive view of ultimate desires, accepting that humans desire all sorts of things ultimately. Schroeder (2004), for example, claims that common sense

[32] In the next chapter, I discuss how Batson (2011) offers an approach that aims to overcome this limitation and test the truth of psychological altruism empirically.

suggests that people have many ultimate desires. People desire "to be well-fed and sheltered and loved, desires that their loved ones be safe, and so on", and, in his view, "[i]n the absence of any reason to override common sense, it seems plausible to agree with it that a desire to be rewarded is only one of many [ultimate] desires a person might have" (Schroeder, 2004, p. 69). For Schroeder (2004), if a desire seems to be ultimate and there is no strong evidence *against* this, there are no reasons to question that it is ultimate (see also Arpaly & Schroeder, 2014).

The easy-access view can account for some phenomena related to ultimate desires. For example, consider the change in the status of desires from instrumental to ultimate (Schroeder, 2020; Sober & Wilson, 1998, p. 221). A good example of this phenomenon is the familiar case of a dad who gives his daughter a cat and ends up loving it. Initially, he might desire the wellbeing of the cat only instrumentally, as a means to his daughter's happiness. A few weeks later, however, the father likes the cat regardless of its effects on his daughter, desiring the wellbeing of the cat ultimately. In principle, there is nothing that precludes us from accepting such a change from instrumental to ultimate[33]. If desires go from instrumental to ultimate, we do not need a sophisticated argument explaining evolutionary causes for our ultimate desires. There might be plenty of other mechanisms forming ultimate desires (see also Stich, 2016, p. 6). The easy-access view is comfortable with the change in the status of desires.

The crucial implication of adopting the easy-access view is that the burden of proof falls upon psychological egoism, for this hypothesis assumes a limited set of ultimate desires. The

[33] A further challenge concerns desires that are *both instrumental and ultimate*. The father's desire to promote the cat's wellbeing might be instrumental to promote his daughter's happiness but also something he desires ultimately — he desires it both instrumentally and ultimately. Thus, the fact that a desire to help others is instrumental to an egoistic desire does not mean that this desire cannot be altruistic as well. That is, altruistic motivation does not need to be "pure" (see Kraut, 2020). A possible response is to say that the father has two different desires, one instrumental and one ultimate, but this is not necessarily the case.

easy-access view allows a very permissive criterion for taking desires as ultimate, and there would be no clear reason why denying that people also want to help others ultimately. So, if we accept the easy-access view, psychological altruism becomes the *prima facie* most plausible hypothesis when compared to psychological egoism (see Schroeder, 2004, p. 5). This would shift the structure of the debate about psychological altruism.

Reading the paragraphs above, a reader might be inclined to see in the easy-access view a key idea to solve the debate about psychological altruism. However, I believe that this view would do more than just solve the debate: adopting the easy-access view would *eliminate* the relevance of the debate itself. By making it trivial to claim that a given desire is ultimate, the easy-access view makes the problem of determining whether ultimate altruistic desires exist a non-problem. The whole debate about psychological altruism erodes and becomes a trivial issue with little relevance. So, if proponents of psychological altruism want to hold that their hypothesis has any sort of relevance, they should be careful in pursuing a defense of the easy-access view uncritically.

This thesis criticizes two central ideas: (1) defining altruistic motivation as an ultimate desire to increase the welfare of others and (2) reducing the broad debate opposing altruistic and egoistic motivations to the debate opposing psychological altruism and psychological egoism. I claim that this is a fundamentally misguided way of thinking about altruism. This chapter discussed several assumptions underlying both (1) and (2), and I hope to have illustrated how these two ideas rely on some fragile assumptions. The easy and hard-access views, in particular, highlight this fragility. It is not my goal here to argue for one of the two views, but since the hard-access view is assumed in the debate, I should state that I do not see any clear argument for

why one should not adopt the easy-access view. This leaves us with an arbitrary assumption at the foundation of the debate about psychological altruism.

The following two chapters are concerned with the main scientifically based cases in favor of psychological altruism[34]. Chapter 4 presents arguments based on research in social psychology, and Chapter 5 presents arguments based on evolutionary psychology. I will argue that adopting the standard account of altruistic motivation makes altruistic motivation inaccessible and produces an unfruitful debate.

[34] As I am interested in analyzing the current debate about psychological altruism, in the next chapters I will follow the literature and assume the hard-access view.

Chapter 4

Our Empathic Nature as the Basis for Psychological Altruism

4.1 Psychological Altruism as an Empirical Problem

The term "empathy" was introduced into the English language in 1909, by psychologist Edward Titchener, as the translation of the German word "*einfühlung*" ("feeling into") (Stueber, 2019). Broadly speaking, empathy refers to the capacity to experience emotions consonant with emotions that others are feeling (Maibom, 2012). Although the term is relatively recent, the idea of empathy is much older, figuring as an important notion in modern philosophy. David Hume (1739/1960) stated that "the minds of men are mirrors to one another" (p. 365), and Adam Smith (1759/2002) considered that the source of our concern for others is our capacity to feel what they are feeling[35] (p. 12). Today, the term "empathy" has been assimilated into both popular and scientific vocabularies. In many areas, empathy is a popular research topic, and this interest has created a complex framework, breaking empathy into several empathic emotions (Maibom, 2009; 2012).

Empathic emotions are considered to play an important role in human cooperative behaviors (de Waal, 2012). However, it is not clear how empathic emotions interact with altruistic motivation. Even if empathic emotions produce helping behaviors, that does not mean that they do so through altruistic motivation. Empathic feelings can cause one to help others for

[35] It should be noticed that both authors called this by "sympathy" rather than "empathy". This use of the term "sympathy" was common in modern philosophy. Today, "sympathy" has a different meaning, which will be discussed in *Section 4.2*.

egoistic reasons. This is an idea with a long history. Darwin (1871/2009) already stated that one might be "impelled to relieve the sufferings of another, in order that our own painful [empathic] feelings may be at the same time relieved" (p. 81). Much before him, Mandeville (1714/1988, Vol. 1) stated that pity "is raised in us, when the Sufferings and Misery of other Creatures make so forcible an Impression upon us, as to make us uneasy" (p. 287).

Since empathy might lead to an egoistic form of help, it is not immediately obvious what role it plays in the debate on psychological altruism. This chapter discusses the idea that empathic emotions can produce altruistic motivation. This is a thesis defended by the social psychologist Daniel Batson, who was one of the first authors to investigate psychological altruism as we conceive it today. Batson's work makes one of the most important cases for psychological altruism in the scientific literature. The vast literature he and his colleagues have produced has influenced and shaped the whole debate on altruism. Batson's research has been regarded by philosophers and psychologists as offering the most robust empirical support for psychological altruism (Stich et al., 2010, p. 169; Sober & Wilson, 1998, p. 260).

Batson has dedicated much of his career to "the altruism question": "[c]ould it be that we are capable of having another person's welfare as an ultimate goal, that not all of our efforts are directed toward looking out for Number One?" (Batson, 1991, p. vii). This is the question of whether psychological altruism is true, and Batson believes to have found evidence supporting a positive answer to it. This chapter explains Batson's case for psychological altruism and discusses to what degree it is successful. Although Batson and colleagues' contribution to the debate on psychological altruism is noteworthy, the conclusion I will argue for is that their case for psychological altruism is not as definitive as he claims. I will discuss some of the empirical and theoretical limitations of his work.

Here is how this chapter is divided. *Section 4.2* explains some of the basic terminology related to empathic emotions, namely, cognitive empathy, affective empathy, emotional contagion, sympathy, empathic concern, and personal distress. Making sense of the complex dimensions of empathic emotions will help us to navigate through Batson's arguments. S*ection 4.3* addresses Batson's theoretical framework, explaining his empathy-altruism hypothesis and discussing the challenges it faces. *Section 4.4* presents the empirical work of Batson, discussing one example in detail. The discussion of this example should give a fair overview of Batson's experimental approach and highlight some of its intrinsic limitations. Finally, in *Section 4.5,* I present an argument directed to Batson's theoretical framework. I argue that there is a crucial problem in how Batson characterizes values and desires. His model only works if we assume that individuals value others' welfare ultimately. But the question of whether we value others' welfare ultimately or instrumentally is subjected to the same difficulties we have to determine whether we desire their welfare ultimately or instrumentally.

4.2 The Empathy Landscape

Since its origins in Titchener (1909/2014), the word "empathy" has been used in different ways. We can see Titchener (1909/2014) already breaking empathy into different meanings, using qualifications such as "motor empathy" and "interpersonal empathy". Ever since research on empathic emotions has produced a complex theoretical framework. The empathic emotions that will be addressed here are *emotional contagion, sympathy, empathic concern, and personal distress*. These emotions interact differently with altruistic and egoistic motivation. To understand Batson's argument for psychological altruism, we need to understand these different

empathic emotions. This section aims to make sense of this framework, providing an overview of the phenomena that fall under the umbrella of empathy.

What is empathy? The first aspect of empathy that we should distinguish is that empathy may be *affective* or *cognitive*. In its cognitive version, empathy concerns taking others' perspectives and understanding their thoughts and feelings (Eisenberg & Miller, 1987). The recent literature on empathy shows that there are two ways of assuming others' perspectives. First, we can imagine what *others* are thinking and feeling in the context in which they are. A second approach is to imagine how *we* would think and feel in others' situations. The first approach is known as *imagine-other*, while the second is known as *imagine-self* (Batson, 2011, p. 18). Distinguishing these two ways of adopting others' perspectives is relevant because they produce different consequences. Batson et al. (1997), for example, have shown that the imagine-self stance is more likely to produce self-centered emotions and behaviors than the imagine-other stance. When imagining ourselves in others' positions, we are ultimately thinking about ourselves. The imagine-other stance, by contrast, pushes us closer to focusing on others and moving beyond ourselves.

The cognitive account of empathy was influential in modern philosophy (e.g., Hume, 1739/1960; Smith, 1759/2002). Recently, however, it is the *affective* account of empathy that has received more attention. We can define affective empathy as follows: "S empathizes with O's experience of emotion E if and only if O feels E, S believes that O feels E, and this causes S to feel E for O" (Sober & Wilson, 1998, p. 234). Affective empathy involves the capacity to recognize others' emotions and to actually *feel* these emotions to some degree. These emotions do not need to be *exactly* the same as that of others: minimally, empathic emotions should be

"consonant — in at least valence (positive/negative), tone, and relative intensity — with those that others experience" (Maibom, 2012, p. 254).

In most cases, empathy is considered a *situational* emotion, being triggered by a certain stimulus. It is possible to think more broadly about empathy as a stable personality trait, making individuals prone to empathize with others. This is the distinction between *dispositional* and *situational* empathy[36] (see Batson, 2011, p. 55). For our purposes here, we can consider empathy situational, as do most authors discussing empathy. Now, let me consider some other empathic emotions that differ from empathy.

A phenomenon slightly different from affective empathy is *emotional contagion*. Similar to empathy, emotional contagion occurs when others' feelings trigger similar feelings in us. However, in emotional contagion, we experience the emotions primarily as our own, without a clear reference to others (Stueber, 2019). As Maibom (2012) explains, "it is easy to become afraid because those around us are afraid, but without thereby being afraid *for* them" (p. 254). Emotional contagion is an ancient trait, and research on primates suggests that these basic empathic emotions may be the main cause of prosocial behaviors in our close ancestors (de Waal, 2006, 2012). The simple empathic mechanism of emotional contagion is present not only in other primates but also in other animals such as pigeons and rats (de Waal, 2008, p. 283).

A good example to illustrate emotional contagion is the phenomenon of *reflexive crying* in newborns. The first controlled study on this phenomenon has shown that 70-hours-old newborns cry when they hear another newborn's cry (Simner, 1971). Further studies also revealed that reflexive crying is a peer-specific behavior: the cry of older babies (11 months old)

[36] The distinction between dispositional and situational is also applied to the debate on altruism. Some authors propose a view of altruism as a disposition, or as a personality trait. I discuss this view in Chapter 9.

and that of chimpanzees do not produce crying in newborns (Martin & Clark, 1982). As Hoffman (1996, p. 65-66) interprets it, these cases show emotional contagion, which can be considered to be a "rudimentary precursor" of empathy. Emotional contagion seems to be produced through a relatively simple pathway that does not rely on a higher-order representational distinction between self and other[37].

Diverging from its meaning in modern philosophy, "sympathy" is defined as an empathic emotional reaction to a perception of a decrease in the welfare of others. "S sympathizes with O precisely when S believes that something bad has happened to O and this causes S to feel bad for O" (Sober & Wilson, 1998, p. 235). This is a reaction to one's perception of others' welfare, so the other is the object of perception. If altruistic motivation is about the welfare of others, this is the empathic emotion that we should look to when discussing psychological altruism. And this is what Batson does. However, he does not use the term "sympathy". Batson (2011) prefers the term "empathic concern", which he defines as an "other-oriented emotion elicited by and congruent with the perceived welfare of someone in need" (p. 11). As Maibom (2012) points out, we can consider Batson's empathic concern to be synonymous with sympathy[38].

Batson explains that empathic concern makes individuals who perceive others in need feel "sympathetic, kind, compassionate, warm, softhearted, tender, empathic, concerned, moved, and touched" (Batson, 2011, p. 103). The key feature distinguishing empathic concern (sympathy) from empathy is that empathic concern is not a reaction to others' *emotions,* but to

[37] A possible neuro-biological basis for emotional contagion could be the well-known *mirror neurons,* which are triggered both when we perform a task and when we observe others performing this same task (Rizzolatti, 2005).

[38] Here I need to highlight an unconventional use of terminology employed by Batson. He uses the term "empathy" to refer to empathic concern. However, this diverges from the standard use of the term empathy, which I explained before.

their *welfare*. A doctor who is about to tell her patient that he was diagnosed with cancer can feel empathic concern for (sympathize with) him, even if the patient is not currently feeling any distress. Once the patient discovers his disease and feels concerned about it, then the doctor can also *empathize* with him. Empathic concern does not depend on others' emotions nor on our capacity to identify them, but on our judgments about their welfare.

The last empathic emotion I will address here is *personal distress*. This empathic emotion occurs when one perceives others in need and such a perception causes aversive feelings. Personal distress is interesting for illustrating how empathic emotions can be linked to egoistic motivation. Just like empathic concern, personal distress is also triggered by the perception that others are in need. It is also a reaction to our perception of others' welfare. However, personal distress is self-centered, while empathic concern is other-centered. When one is in personal distress, the object of one's perception is one's aversive feelings, and one's goal is to relieve one's own distress, regardless of what happens to others.

The list of aversive states that one can feel when in personal distress includes "alarmed, bothered, disturbed, upset, troubled, worried, anxious, uneasy, grieved, and distressed" (Batson, 2011, p. 103). In Bloom (2018) we find a case that illustrates very well the essence of personal distress. The case is set in World War II and is about a woman who, living near concentration camps and often seeing the atrocities committed against the prisoners, wrote a letter to the Nazi officials complaining:

> One is often an unwilling witness to such outrages. I am anyway sickly and such a sight makes such a demand on my nerves that in the long run I cannot bear this. I request that it be arranged that such inhuman deeds be discontinued, or else be done where one does not see it. (Bloom, 2018, p. 74)

What bothered her was not the fact that people were suffering, but the aversive emotional reaction that their suffering produced in her. In empathic concern, we care about others, about what happens to them, while in personal distress the focus is on relieving the aversive arousal that the perception of others in need causes us. As Batson (2011) states, the importance of distinguishing between personal distress and empathic concern "is underscored by evidence that parents at high risk of abusing a child are the ones who more frequently report distress at seeing an infant cry" (p. 19).

This section introduced a rich repertoire of empathic emotions. These emotions are similar, and distinguishing them is not an easy task. The most common means to access empathic emotions is the use of self-reports and questionnaires. Batson's experiments, for example, rely heavily on these instruments. However, there are important limitations to these instruments. For example, Eisenberg et al. (1988) comment that "there are reasons to believe that both questionnaires and self-reports… are sometimes affected by the desire to be perceived positively by others as well as by self-deception" (p. 766). The limited access to one's own mental states, the possibility of confabulating, and the social expectations, all go against the credibility of self-reports (see Maibom, 2012). Individuals might state what they want to be true rather than what is true; individuals might be wrong about their own subjective states; individuals might say what they believe others want to hear; and so on[39].

An alternative to self-reports and questionnaires is the use of physiological measurements. For example, heart rate has been shown to indicate sympathy, personal distress,

[39] Batson (2011, p. 56) criticizes self-reports in the case of dispositional empathy, pointing out the problems I mention here. It is true that reports about one's dispositions are more subject to these problems, but it seems plausible that the same criticism can be applied to measurements of situational empathic responses.

and empathy by different studies — and the same goes for skin conductance and startle reflex (Maibom, 2012, p. 256). However, since the responses measured can often be triggered for different reasons, the results are difficult to interpret. The problem, in short, is that physiological measurements are not as precise as we need in order to distinguish similar phenomena such as personal distress and empathic concern. This difficulty is particularly pronounced when we consider that these empathic responses are likely to occur together to some degree. "Physiological measures, therefore, do show interesting correlations with subjects being exposed to others in distress, but are not very precise indicators of what emotions, exactly, people experience" (Maibom, 2012, p. 256).

There is a rich framework surrounding the idea of empathy. The different empathic emotions discussed here are all important pieces in the mechanism allowing humans to cooperate with each other. But coming back to the debate on psychological altruism, how can we make sense of these empathic emotions in the context of the debate on altruistic motivation? More precisely, what is the causal relationship between empathic emotions and ultimate altruistic desires? The next section discusses Batson's hypothesis, which claims that one of the empathic emotions, empathic concern, can *produce* altruistic motivation.

4.3 The Empathy-Altruism Hypothesis

Altruism was not very popular among the most influential psychologists from the past. Authors such as Sigmund Freud and William James endorsed egoistic accounts of human psychology, while others, such as B. F. Skinner, rejected the whole debate about egoistic or altruistic motivations (Batson, 1991, p. 35-39). In the early 1900s, "theories of motivation based

on behaviorism or psychoanalysis were sufficiently sophisticated to provide an egoistic account of any behavior that might appear to be altruistically motivated" (Batson, 1987, p. 66). In the 1960s, however, research on prosocial behaviors became gradually more popular in social psychology, and in the 1970's the question regarding whether the *motivations* for these behaviors are altruistic started to be considered (Batson, 1991, p. 42).

In this early rise of altruistic motivation in social psychology, altruism was not taken to be an ultimate desire to benefit others. As Batson (1991) comments, in one of these early accounts of altruism, the opposition between altruism and egoism was conceived in terms of the opposition between *internal* and *external rewards* (p. 44). Cialdini and colleagues, for example, have argued that altruism is explained through internalized self-rewards, which we acquire through education (Cialdini & Kenrick; 1976; Cialdini et al., 1981). What Cialdini and colleagues call "altruism" is considered nothing but a special form of egoism in the contemporary approach to altruism (Batson, 1991, p. 57). Batson criticized this way of conceiving altruism. In his view, these were still egoistic forms of helping, not a genuine form of altruism. Looking for a more substantial account of altruism, Batson (1991) dismissed these early accounts of altruism, accusing them of being "pseudo-altruistic" (p. 43).

Dissatisfied with the accounts of altruism proposed by his colleagues, Batson started to defend that the proper way of conceiving altruism was in terms of ultimate desires to increase the welfare of others. Batson's book *The Altruism Question* (1991) is a milestone in the integration of the standard account of altruistic motivation into social psychology. The altruism question, with Batson's influence, became the question of whether humans have ultimate desires to help others. The conclusion that Batson reaches is that the evidence from the experiments he conducted throughout the years supports the thesis that altruistic motivation exists. This section

explains Batson's theory and the challenges it faces. In the next section, I will discuss the empirical approach conducted by Batson.

A starting point of Batson's theory is the existence of substantial evidence showing that empathic concern increases the likelihood of producing helping behaviors[40] (Batson, 2011, p. 70). Batson proposes that such helping is caused by altruistic motivation. This is what Batson (2011) calls the *empathy-altruism hypothesis*, which states that "[e]mpathic concern produces altruistic motivation" (p. 11).

Although the helping behavior correlated with empathic concern may be based on altruistic motivation, there is nothing precluding it from being based on egoistic motivation. The helping behavior linked to empathic concern is no evidence that the motivation is altruistic (Stich et al., 2010, p. 175). So, in order to support his empathy-altruism hypothesis, Batson needs to dismiss the competing egoistic alternatives. But this is no easy task. As he explains, the empathy-altruism hypothesis cannot be tested directly, but only indirectly: it is only when we fail to explain the data with the egoistic hypothesis that we can find support for the empathy-altruism hypothesis (Batson, 2011, p. 106). Let me present the main egoistic hypotheses addressed by Batson.

Three main egoistic hypotheses are competing against the empathy-altruism hypothesis to explain the help in empathic concern: *empathy-specific rewards*, *empathy-specific punishment avoiding*, and *aversive-arousal reduction* (see Batson, 2011, pp. 70-75). Each of these

[40] Note, however, that this is not exclusive to empathic concern. Personal distress is also likely to increase helping, although it does so as a means to reduce the aversive stimulus. Empathy is also linked with increasing helping behaviors, both in its cognitive and affective versions (Oswald, 1996).

hypotheses will make distinct behavioral predictions. So, the test of the hypotheses is based on the observation of behaviors. Now, let me explain briefly the three egoistic hypotheses[41].

Consider the *empathy-specific reward* hypothesis, first. Some egoistic individuals help others only because they believe that helping will bring them some sort of reward (Batson, 2011, p. 61). These rewards can be external, such as social praise, or internal, such as the good feeling associated with believing to be a good person. The empathy-specific rewards hypothesis claims that the increase in helping presented by individuals under empathic concern can be explained by individuals' expectations of rewards (Batson, 2011, p. 70). For example, one can associate higher rewards with cases in which one feels empathic concern, or one might learn that social rewards are often higher in cases where we feel empathic concern. Empathic concern could be some sort of signal for egoists, indicating a possible reward, and this could explain why individuals feeling empathic concern are more likely to help others.

How is one to know whether a helping behavior is based on reward-seeking? Batson postulates that, since the rewards of helping go for the helper, egoists seeking rewards of helping are likely to try to be the one who helps. Having someone else to help would not bring about the rewards produced by helping. So, we can expect that reward-seekers will either help directly or not help at all. This differs from individuals moved by altruistic motivation, who will not be concerned with who is helping as long as the person in need is helped.

The difference between helping and having someone helping the person in distress is a clear, observable behavioral difference based on which we can test these two hypotheses and see

[41] The list of three egoistic hypothesis discussed here is not exhaustive. Batson actually proposes *six* egoistic hypotheses (Batson, 2011, p. 76). However, these six are only variations of the three main hypotheses covered here (Batson, 2011, p. 106). Since the e aversive-arousal reduction hypothesis is still the most relevant of them, this is the one I will give more attention, discussing it more in the next section.

which one better explains the data. Another related behavioral prediction is that the reward seekers will not care whether helping is *effective* (Batson, 1991, p. 102). All they need is to present some sort of virtue signaling, where others perceive them as helping or trying to help. They will not bother if the helping turns out to be inefficient, "as long as the ineffectiveness is justified" (Batson, 2011, p. 75). Altruists, by contrast, will be concerned about the effectiveness of helping, since their goal is to help the person in need.

Consider now the *empathy-specific punishment avoiding* hypothesis (Batson, 2011, p. 72). This hypothesis states that individuals help as a means to avoid the punishment that follows from not helping someone to whom they feel empathic concern. These punishments can be either *external* (e.g., social sanctions) or *internal* (e.g., guilt or shame). The behavioral predictions of this hypothesis are different from the previous egoistic hypothesis. For egoists avoiding punishment, differently from the reward-seeking case, it is preferable if others do the helping behavior. Since individuals avoiding punishment are only trying to avoid being held accountable for *not* helping, then if someone else helps the person in need that will be less costly for the punishment avoider (Batson, 1991, p. 102).

Punishment avoiders, just like individuals moved by altruistic motivation, will not be concerned with who is helping. This means that we cannot rely on this behavioral variable to distinguish punishment-avoiders from altruists. However, a different variable can be used. Egoists whose motivation aims to avoid punishment will not care so much about whether helping is effective. All they care about is signaling that they are helping, not whether the helping is effective. Therefore, the *empathy-specific punishment avoiding* hypothesis also has a clear behavioral prediction that one can use to distinguish it from altruistic motivation. So, when it

comes to whether helping is effective, punishment avoidance will predict the same behaviors from reward seekers. Namely,

Finally, consider the egoistic hypothesis that is both the most common and most challenging, namely, the *aversive-arousal reduction* hypothesis. This hypothesis says that individuals feeling empathic concern help others in order to reduce the feeling of empathic concern, which they experience as aversive (Batson, 2011, p. 73). The aversive-arousal hypothesis claims that individuals are helping others simply as a means to reduce the aversive arousal of empathic concern, not out of a selfless concern for others — "[t]hey help for the same reason you turn down the thermostat when the room is too hot" (Sober & Wilson, 1998, p. 261).

The aversive-arousal reduction hypothesis represents the main challenge for the empathy-altruism hypothesis. This is so because both hypotheses predict almost the same behaviors. Firstly, both hypotheses predict that individuals will not care about *who* is helping, as long as the needs of others are reduced. Secondly, for both hypotheses individuals will care about the effectiveness of helping, since arousal only ends when individuals believe the person is no longer suffering. The two behavioral differences that could distinguish egoists and altruists in the first two egoistic hypotheses are not helpful in the case of the aversive-arousal reduction hypothesis.

Although the aversive-arousal reduction and the empathy-altruism hypothesis have similar behavioral predictions, Batson believes that we can still distinguish their behavioral manifestations. He argues that, differently from altruists, aversive-arousal egoists will prefer to *escape* from the aversive arousal if that is an option. "[I]f helping is at least moderately costly, and empathically aroused individuals can easily escape continued exposure to the person's suffering, they [the egoists] will do that instead of helping" (Batson, 1991, p. 104). The next section discusses in more detail Batson's use of the easy escape.

4.4 The Experimental Approach

In this section, I will briefly discuss Batson's experimental work and ask whether it offers substantial evidence for psychological altruism. My discussion here will not be as complete as Batson's extensive work deserves. But rather than superficially surveying several of these studies, my strategy will be to explore one experiment in detail. The experiment I will discuss (Batson et al., 1981) gives a glimpse of the main techniques used in Batson's experiments. This experiment aims to reject the *aversive-arousal reduction hypothesis*, which is the most challenging egoistic hypothesis due to the similarity between its behavioral predictions and the predictions of the empathy-altruism hypothesis. This section illustrates the kind of approach adopted by social psychologists investigating psychological altruism, highlighting both the strong and weak points of such methodology.

In a study with 44 female psychology students[42], participants watched, through closed-circuit television, a woman who they were told to be another participant (Batson et al., 1981). Subjects were told that the woman they are watching (named "Elaine") will receive ten electric shocks. The video shows Elaine indicating, through facial and bodily expressions, that she finds the shocks very painful. After watching Elaine receive two shocks, the subjects are offered the

[42] Stich (2016) discusses the implications of the samples used in Batson's work. What is claimed to be an overview of human nature is, in fact, a snapshot of a very select group: people from "WEIRD" societies (Western, Educated, Industrialized, Rich, and Democratic), particularly living in the USA and studying psychology, who are common in Batson's studies. This criticism is important for highlighting one of the limitations of Batson's work, and Batson himself is aware of this problem (Batson, 2011, p. 106). However, this criticism does not undermine Batson's conclusions. That is because psychological altruism is an existential claim. Even if it turns out that American psychologists are the only altruists in the human species, psychological altruism would still be true.

chance of switching places with her, receiving the shocks in her place. Deciding to switch places with Elaine is the helping behavior that the experiment aims to study.

Two variables are manipulated in this study. The first is the *degree of empathic concern* that participants are likely to feel for Elaine. There are different techniques to manipulate high and low empathic concern. One of them, which is used in Batson et al. (1981), is presenting subjects with different profiles representing Elaine. First, subjects fill out a questionnaire informing their values and interests. After this, they receive the same questionnaire with what they were told to be Elaine's answers. In the *low* empathic concern group, Elaine's responses are radically different from that of the participants, while in the *high* empathic concern group, Elaine is presented as sharing the participants' values and interests. This manipulation is based on the assumption that individuals tend to empathize more with those with whom they share common interests and values, and is a common practice in social psychology[43] (Batson, 1991, p. 114; Stich et al., 2010, p. 174). It is not assumed, of course, that *all* individuals in the high empathic concern group will empathize more, but that the *average* level of empathic concern will be higher in this group (Sober & Wilson, 1998, p. 262). As I said before, individuals feeling empathic concern are more likely to help. But this is a prediction shared by both the egoistic and altruistic hypotheses. So, we need another variable to test them.

The second variable manipulated is the presence of an *easy escape*. As I said, subjects believe that Elaine will go through ten shocks. Participants in the *easy* escape condition, however, were told that they will only watch her receiving *two* shocks. By contrast, participants in the *difficult* escape condition were told that they will have to watch Elaine going through all

[43] Batson et al. (1981) also use another method called "emotion-specific-misattribution technique" (see Batson, 2011, p. 101). The idea is to suggest to subjects that their emotional reaction is a consequence of a placebo.

ten shocks. Both groups know that Elaine will receive ten shocks, so what is different is that in the easy escape condition, participants will not have the distress of watching her suffer (Batson, 2011, p. 99). The prediction proposed by Batson is that if individuals are helping only to relieve their personal distress caused by the perception of Elaine's distress, then they will be less likely to help when escaping is an option.

Given the two variables above, there are four experimental settings: high empathic concern with easy escape; high empathic concern with difficult escape; low empathic concern with easy escape; and low empathic concern with difficult escape. The egoistic hypothesis tested in this experiment is the aversive-arousal reduction hypothesis. Both this egoistic hypothesis and the empathy-altruism hypothesis predict that individuals with higher empathic concern help more than individuals with low empathic concern. The different predictions based on which this experiment can test the two hypotheses concerning the easy escape scenario. The empathy-altruism hypothesis predicts that the helping of the individuals in the high empathic concern group *will not be reduced* in the easy escape condition. By contrast, the aversive-arousal reduction hypothesis predicts that the high empathic concern group will *be less likely to help* when there is an easy escape. This is the difference being tested in the experiment.

The results of this experiment support the empathy-altruism hypothesis. Individuals with high empathic concern were not less likely to help in the easy escape condition. Ten similar experiments were conducted, reaching similar results: helping in high empathic concern participants was not significantly reduced by the presence of an easy escape (Batson, 2011, p. 112). Therefore, if we accept the predictions attributed to the hypotheses, at least sometimes helping cannot be explained by the aversive-arousal reduction hypothesis.

Some authors have remained skeptical regarding the effectiveness of the easy escape condition. In the experiment discussed, the easy escape consists of not watching Elaine receiving shocks. However, it is not clear whether leaving the room *knowing* that Elaine will still receive shocks should be considered an easy escape for subjects in the high-empathy category. As Sober and Wilson (1998) state, it is possible that individuals in the high empathic concern group "realized that they would retain painful memories of the needy other if they declined to help" (p. 264). If this is the case, then leaving the room would *not* constitute a true psychological escape. Wallach and Wallach (1991) claim that "[t]o offer escape from witnessing the suffering of the victim, therefore, does not offer removal of the aversive arousal produced by that suffering" (p. 153).

Batson (2011) responds to this criticism, claiming that evidence suggests that easy escape provides psychological escape (p. 137). But Batson (2011) recognizes that this is not conclusive: "physical escape *apparently* [emphasis added] does provide psychological escape" (p. 139). This highlights one important aspect of Batson's case for psychological altruism. The evidence he provides is not a definitive blow against psychological egoism. As I said earlier, the test of the empathy-altruism hypothesis is indirect: "we can accept the empathy-altruism hypothesis as valid only as long as no new plausible egoistic explanation can be proposed that accounts for the existing evidence" (Batson, 2011, p. 106). As Batson (2011) recognizes, even if experiments indicate that none of the egoistic hypotheses available explain the data, "we can never be sure that we have exhausted the set of all plausible alternative explanations" (p. 106).

Another criticism of Batson's theory is raised by Stich et al. (2010), who claim that Batson's experiments only address a specific kind of instrumental desires, not considering cases in which instrumental desires are "long-standing" (p. 195). Long-standing instrumental desires

are instrumental desires that acquired certain stability in one's decision-making. To illustrate this kind of desire, Stich et al. (2010) mention the example of a desire to pay a monthly electric bill. The desire to pay the bill is an instrumental desire, since we only want to do it in order to get the benefits that follow from having electricity. But this desire to pay the bill can achieve a degree of stability in our decision-making so that we do not have to revisit the reasoning behind it when we activate such a desire. We see the bill and activate the desire to pay for it, not thinking about the reasons why we want to do so. Now, think about the example of altruistic desires. It is possible that one's desire for avoiding punishment, such as feeling guilt, for example, produced long-standing instrumental desires to help others in need. When we see someone in need, we only experience the desire to help, as if it were ultimate, although, in reality, it is merely instrumental. I believe that the criticism raised by Stich et al. (2010) can be further developed if we take a closer look at the idea of "activating" desires in Batson's framework.

Remember that Batson's empathy-altruism hypothesis claims that empathic concern *produces* altruistic motivation. It seems that the empathy-altruism hypothesis is about the *creation* of altruistic motivation: empathic concern seems to be part of a mechanism that could explain how altruistic motivation is created. However, this is not necessarily the case. When discussing the ways in which empathic concern may produce altruistic motivation, Batson (1991) says that one of the effects of empathic concern is to change the magnitude of an already present altruistic motivation (p. 87). Such a change in the magnitude or intensity of desires can "*activate*" these desires (Batson et al., 1995, p. 300). That is, one might have an altruistic desire that is not very intense and would not produce action, but once one feels empathic concern, this desire becomes stronger and can produce behavior.

The idea of empathic concern activating desires to help seems plausible. If I *already* desire ultimately to increase the welfare of my friend, it seems reasonable to expect that feeling empathic concern for him is likely to increase the intensity of this desire, perhaps triggering a desire that would otherwise remain dormant. But should we call this a "production" of altruistic motivation? Batson seems to think that we should. In a passage, Batson (2011) says that "[e]mpathic concern activates the desire to reach the goal of eliminating the perceived need of the person for whom the empathy is felt. *That is* [emphasis added], empathic concern produces altruistic motivation" (p. 31). Batson equates "activation" with "production". This, however, raises questions about to what degree the evidence collected by Batson is supporting psychological altruism. To better understand the issue, we need to take a deeper look at the egoistic hypotheses addressed by Batson.

The egoistic hypotheses considered by Batson aim for an egoistic goal (a reward, a reduction of aversive arousal, etc.) and use altruistic *helping* as a means to get these goals. In this case, we can offer alternative means of getting the egoistic goal and individuals are likely to choose the easier way to get it. This is the assumption of Batson's experiments. But there is a more complex way in which our desires to help others can be instrumental. Imagine, for example, that I realize that a life worth living is a life where I value others' welfare regardless of the immediate benefits that they cause me. If I form this desire to help others, based on my ultimate (egoistic) desire to have a life worth living, then my desire to help others will be instrumental. But this instrumental desire to help others can become stable, like the long-standing desires discussed by Stich et al. (2010). My behavior will be considered altruistic in Batson's experiment, for I will not trade this important desire to help others for a mild social reward or to reduce aversive arousal. It is not clear what experimental conditions would be able

to test such a strong and stable (but instrumental) desire to help others. Empathic concern would activate this desire to help, but this desire is instrumental. Batson does not provide an argument for why the desires to help others that are activated by empathic concern are necessarily ultimate.

So, we can raise the following challenge for Batson's theory: if empathic concern is merely *activating* our desires to help others, then why not postulate that they are activating only these stable, long-standing instrumental desires to help others? Batson's experiments rule out egoists that see helping behavior as a means to get to a goal, but it does not satisfactorily rule out egoists who have instrumental stable desires to increase the welfare of others. In short, it excludes *instrumental helping*, but it does not exclude *instrumental desiring*. That is, it rules out egoistic desires that use helping as a means to obtain something, but it does not exclude desires to help others that are instrumental to egoistic desires.

Before ending this section, I will address a criticism that can be raised against my treatment of Batson's views. One might complain that I have set the bar unfairly high for psychological altruism. If we have two competing hypotheses, we should accept the one for which we have more evidence. So, one could say that even if the evidence we have to believe in psychological altruism is very weak, this is still a good case for it.

In order to respond to this criticism, I want to remind the reader of the last section of Chapter 2. There, I discussed why egoistic explanations are considered to be *prima facie* more plausible than altruistic explanations, giving psychological egoism an initial advantage when compared to psychological altruism. Psychological egoism is simpler, it assumes only the existence of desires that are already uncontroversial, and it is more easily explainable through natural selection. More importantly, as I mentioned there, this initial advantage of egoistic motivation is accepted by authors such as Batson (1991, p. 50). So, in adopting egoistic

explanations whenever they are available, I am just following the method adopted by Batson himself. The evidence provided by Batson, in my view, is just not strong enough to overcome this initial advantage of psychological egoism. I will argue that the same is true for the evolutionary arguments, which I will discuss in the next chapter.

This section discussed only a sample of the rich experimental work of Batson and colleagues and the rich discussion surrounding it, highlighting some limitations of Batson's approach to altruistic motivation. I pointed out that it is hard for Batson to distinguish ultimate altruistic desires and long-standing, stable instrumental desires to increase the welfare of others. The next section takes a closer look at the process of activation of altruistic motivation as it is understood in the empathy-altruism hypothesis.

4.5 The Problem of Ultimate Values

In Batson's theory, there are a few conditions that need to be in place in order for empathic concern to "produce" (activate) desires to help others. In his early work, Batson considered that these conditions were simply (1) perceiving someone in need and (2) adopting the others' perspective (Batson & Shaw, 1991, p. 112; Batson, 1987, p. 91). In his more recent work, however, Batson changed his view on the issue. He kept the first condition but dropped the second[44]. He explains that this is because one can feel empathic concern without having to adopt others' perspectives and can also adopt their perspective without feeling empathic concern (Batson, 2018, p. 195). In short, perspective-taking is *neither necessary* nor *sufficient* for empathic concern to elicit altruistic motivation.

[44] For a criticism of condition (2), see Stich et al. (2010, p. 174).

In his late work, Batson (2011) considers that the two conditions for empathic concern to produce altruistic motivation are (1) perceiving someone in need and (2a) *to value ultimately this person's welfare* (p. 44; see also Batson, 2018, p. 189). It is this last condition that will be analyzed in this section. There are two separate assumptions underlying condition (2a). First, it is assumed that we can know whether a value is *ultimate* or *instrumental*. Second, it is assumed that we can distinguish *desires* from *values*. I will argue that neither of these assumptions is well supported.

Consider the first one. Underlying condition (2a), there is the assumption that we can know whether individuals are valuing others' welfare ultimately or instrumentally. This is so because Batson (2011) establishes that the valuing of others' welfare needs to be *ultimate,* not instrumental (p. 45; see also Batson, 2018, p. 197). But how can Batson tell whether the valuation of others' welfare is ultimate or merely instrumental?

To access subjects' values, Batson relies on *self-reports* and *questionnaires* (Batson et al., 1995). These instruments are expected to tell whether subjects value the welfare of the person introduced to them in the experiments. Not only that but these self-reports and questionnaires are expected to tell whether subjects are valuing others' welfare *ultimately* or not. The following is an example of the kind of question that subjects in the experiments find:

> "How much do you value this person's welfare?" with responses ranging from 1 (*somewhat*) to 9 (*very much*); "How important is it to you that this person is happy?" with responses ranging from 1 (*not at all important*) to 9 (*very important*); and "How important is it to you that this person does not suffer?" (same scale as previous item). (Batson et al., 1995, p. 302)

In another experiment, the question formulated to assess one's valuation of another's welfare asks:

> Think of someone whose welfare you value very highly (e.g., a best friend or favorite family member). Now think of someone whose welfare you do not value highly (e.g., someone you know nothing about at all). Compared to these two extremes, how much do you value the welfare of Participant ____? (1 = not highly (like a person you know nothing about), 9 = very highly (like a best friend)). (Batson et al., 1995, p. 306)

As discussed in *Section 4.2*, the answers to these questions will not always reflect the subjects' true mental states. Introspection and the observation of behaviors are fallible, for values can be instrumental to other values and we might not be aware of that. I might believe that I value my friend's welfare ultimately when, in reality, I value it instrumentally, for I feel good in his company, for example. In the same way that we may not have conscious access to desires, we often will not have access to our values and the causal chain linking ultimate values to instrumental values. There are significant challenges precluding us from knowing whether values are ultimate through self-reports and questionnaires.

Considering the discussions in the previous chapters, we can see the similarity between the problem above and the *deadlock problem*. This later problem, remember, concerns the challenge of knowing whether a desire to help is ultimate or instrumental. My view is that, when trying to identify whether values are ultimate, we should expect *the same* difficulties that we have in identifying whether a desire is ultimate. We have the deadlock problem all over again, but now, instead of a problem about desires, it is about values. The two versions of the deadlock problem present the same epistemic challenges and deserve the same degree of skepticism.

Although I believe that there are reasons to be skeptical about the reliability of self-reports and questionnaires, my criticism of Batson will not rely on dismissing the method of using questionnaires and self-reports. My argument here is much simpler. I argue that there is an *inconsistency* in Batson's approach to values and desires. Batson's quest to answer whether our

desires to help are ultimate assumes that it is very hard for us to know whether a desire to help is ultimate or merely instrumental. All his empirical work follows from this challenge. But it is precisely *because* Batson has such a cautious and skeptical stance regarding ultimate desires that he is not justified in not being equally cautious and skeptical regarding ultimate values. If self-reports and questionnaires are not reliable means to know whether *desires* are ultimate, then they cannot be considered reliable means to know whether a *value* is ultimate.

Let me now address the second assumption underlying condition (2a). This assumption concerns the distinction between desires and values. If Batson's theory takes the existence of ultimate values as a condition for the production of ultimate desires, then we need to understand the difference between values and desires. I will argue that Batson does not provide a satisfactory distinction between these two concepts.

Although Batson usually refers to altruistic motivation using "motivational state" or simply "goal", he has been using the term "desire" to refer to altruistic motivation since his early work (see Batson, 1991, p. 230). The central role of desires becomes much clearer later, in his *Altruism in Humans* (2011), in which Batson presents a more sophisticated theoretical framework when compared with his early work. Already in the introduction of this work, Batson defines altruism as "a desire to benefit someone else for his or her sake rather than one's own" (Batson, 2011, p. 3). Later, Batson explains in more detail what is the motivational state that he identifies as altruism. He claims that this motivational state is a "goal-directed force", which is composed of four features:

> (a) The individual desires some imagined change in the experienced world (neither the desire nor the imagined change need be conscious). This is what is meant by a goal. (b) A force of some magnitude exists, drawing the individual toward the goal. (c) If a barrier prevents direct access to the goal, alternative

> routes will be sought. (d) The force disappears when the goal is reached. (Batson, 2011, p. 20-21)

Batson's account of desire is not different from the standard account of altruistic motivation discussed in the previous chapters. In his more recent book, Batson (2018) reiterates the view that altruistic motivation is a desire (p. 22). Thus, despite the fact that "desire" is a word not used very often by him, desires constitute an important element in Batson's framework.

But while his account of desire is clear, the same is not true for his account of value. Batson does not provide a clear definition of values and does not discuss the difference between desires and values[45]. A remarkable indication that Batson does not consider the importance of the difference between desires and values is that, in some passages, he uses desires and values interchangeably (e.g., Batson, 2011, pp. 30-31). But, if there is a causal relationship between values and desires, where the former is assumed in order for us to have the latter, we need to understand what is meant by "value".

Defining "value" is no easy task. Value theory is a complex topic in philosophy, and even among naturalistic accounts of value, there is much disagreement about how to characterize values (see Mason, 2018; Schroeder, 2021). Our understanding of values will affect important issues, such as the debate about which normative theory is correct. In this section, I will not discuss the philosophical problem of defining values, nor argue for which account of value is the best. Here, I am only interested in understanding the meaning of value in Batson's framework.

Since Batson does not provide a clear definition of values, one way of understanding his account of value is to look at some of the works he references. When distinguishing instrumental

[45] He does distinguish, however, values from *emotions*, claiming that values are more stable dispositions than emotions, not depending on the occurrent perception of others' needs (Batson et al., 1995, p. 301).

and terminal (ultimate) values, Batson (2011) draws on the work of Rokeach (1973) (p. 45). Rokeach's account of value might shed some light on Batson's views on values[46]. Rokeach (1973) defines a value as "an enduring belief that a specific mode of conduct or end-state of existence is personally or socially preferable to an opposite or converse mode of conduct or end-state of existence" (p. 5). He explains that what he is calling a "belief" in this definition is a "prescriptive belief", "wherein some means or end of action is judged to be desirable or undesirable" (1973, p. 7). Values, in his view, are cognitions about what is desirable and are felt positively or negatively. More importantly, Rokeach (1973) also claims values *motivate* actions, claiming that "a value has a behavioral component in the sense that it is an intervening variable that leads to action when activated" (p. 7).

The problem should be clear: Batson's framework relies on the distinction between values and desires. However, if he adopts an account of desires similar to that of Rokeach, it is not clear how one can distinguish values from desires. Rokeach's (1973) account of values fits nicely into the action-based theory of desire, which states that "[t]o desire that P is to have a mental representation that P which plays a certain causal role, namely, that of disposing one to bring it about that P" (Schroeder, 2004, p. 24). The same action-based theory also represents Batson's account of *desire*. So, it is not clear how we should distinguish desires from values in Batson's framework[47].

[46] Batson (2018) also refers to the work of Hepach, Vaish, and Tomasello (2013), claiming that they provide evidence that "the capacity to value (care about) another's welfare emerges somewhere between 1 and 3 years of age" (p. 194). However, the authors only talk about "care about", not about "value". So, it seems that Batson is using the term value in a loose way, which, as I will argue, is problematic.

[47] Notice that my criticism is not directed to Rokeach's account of value, but to Batson's use of this account in his framework. As Gaus (1990) points out, the idea that values guide actions is one of the ways in which values are commonly understood in philosophy.

Batson's answer to the altruism question relies on the assumption that people value others' welfare ultimately. But if Batson assumes the existence of these values as a condition for the existence of ultimate desires to increase the welfare of others, and if he does not offer a clear way of conceiving the difference between these two states, then we have a problem. The point I defended in this section is that Batson is solving one problem by creating another one with the exact *same explanatory deficiencies*. One can ask the "altruism question" in terms of values: are our values ultimately altruistic or egoistic? This would produce a debate identical to the debate on whether desires are ultimately altruistic or egoistic. If values and desires are identical in Batson's framework, then we can say that Batson's theory commits a *petitio principii*, assuming what it is trying to argue. More charitably, we can say that, until Batson clarifies this distinction, it is not clear whether his work responds satisfactorily to the altruism question.

The problem presented here is not exclusive to Batson's framework. The proximity between desires and values is a common problem in philosophy. Preference satisfaction views of values, for example, claim that values are, ultimately, desire satisfaction (Mason, 2018). In this view, valuing x is desiring x. In theories of desire, the line dividing desires and values is also often unclear. This is particularly clear in the "good-based theories", which claim that to desire x is to value x (Schroeder, 2020). Thus, the theoretical limitation of Batson's framework follows not from a simple mistake, but from a complex philosophical issue.

In this chapter, I discussed the work of Batson and how empathic emotions can be linked to altruistic motivation. Despite the criticism raised against Batson's work, it is not my intention to dismiss the value of his fascinating empirical work. Batson rejects some simplistic egoistic hypotheses, offering genuine progress in the debate on psychological altruism. However, a careful analysis of his theory raises some concerns. It is not clear to what degree his work

supports the claim that the altruistic motivation considered in the empathy-altruism hypothesis is really an ultimate desire to increase the welfare of others. His experiments do not exclude cases in which instrumental desires to increase the welfare of others are stable and long-standing. This last section added a further concern regarding the assumption of ultimate values, which seems to defeat the strength of his response to the altruism question. After all, if Batson's response assumes that people value others' welfare ultimately, then the question is: is this valuing ultimate or instrumental? The altruism question would be simply stated in terms of values rather than desires. In the next chapter, I move on to the second most influential case for psychological altruism in the scientific literature, namely, the evolutionary case, where authors claim that altruistic motivation is adaptive.

Chapter 5

The Evolution of Altruistic Motivation

5.1 From Behaviors to Evolutionary Causes

Empirical work in social psychology has advanced the debate over psychological altruism, distancing it from its past as an intractable aprioristic endeavor. But the idea of bringing the debate closer to a scientific approach is not exclusive to social psychologists. In this chapter, I discuss another argument for psychological altruism from the scientific literature. This is the evolutionary case for psychological altruism, which tries to respond to the altruism question by arguing that altruistic motivation is *adaptive*. This chapter aims to present the evolutionary arguments for psychological altruism and discuss some problems that they face.

The evolutionary argument for psychological altruism was originally proposed in the book *Unto Others* (1998), by the philosopher Elliott Sober and the evolutionary biologist David Sloan Wilson. After defending that the experimental approach from social psychologists was inconclusive, Sober and Wilson (1998) propose a different strategy to address the altruism question: “to shift the focus from behavioral effects to evolutionary causes" (p. 298). The evolutionary approach to the debate on psychological altruism has inaugurated a new way of thinking about psychological altruism, inspiring much debate (Batson, 2000; Clavien, 2012; Harman, 2000; Jamieson, 2002; Kitcher, 2011; Lemos, 2004, 2008; Piccinini & Schulz, 2018, 2019; Rottschaefer, 2000; Schulz 2011, 2016, 2018; Stich, 2007, 2016; Stich et al., 2010).

In short, the idea of the evolutionary case for psychological altruism is to show that, under certain evolutionary conditions, altruistic motivation is likely to be selected. This argument

is based on the principles of evolutionary psychology, which views psychological features as phenotypes subject to natural selection as much as any other phenotype (see Buller, 2005). In this approach, altruistic and egoistic ultimate desires are *psychological traits* subjected to natural selection. In order to determine which of these traits is more likely to evolve, we need to be able to compare them from an evolutionary point of view.

However, psychological traits are selected based on the behaviors they produce. So, in the case of altruistic and egoistic motivation, we have a problem: as explained before, both motivations can produce the same *kind* of helping behaviors. Egoistic motivation can produce all sorts of instrumental desires, which will produce the same helping behaviors as that produced by altruistic motivation. So, the kind of helping behavior itself cannot be the criterion based on which we can compare these motivations. We need another variable to indicate a different selective pressure over these two motivations.

The ingenious strategy of Sober and Wilson (1998) is to compare altruistic motivation and egoistic motivation not in terms of the kind of helping behavior they produce, but in terms of their differences in *triggering* helping behavior. They argue that altruistic and egoistic motivations differ regarding the circumstances in which they trigger helping behaviors. As Sober (1994) says, "[w]henever natural selection causes a behavior to evolve, it must equip the organism with a mechanism that triggers the behavior in the appropriate circumstances" (p. 17). Even if altruistic and egoistic motivations can produce the same kinds of behaviors, they will trigger these behaviors differently, so "it remains possible that one of them was more likely than the other to have evolved" (Sober & Wilson, 1998, p. 298). Altruistic and egoistic motivations are what evolutionary psychologists call "proximate mechanisms*"* competing to control helping behavior (Buller, 2005, p. 51).

In this chapter, the version of psychological egoism discussed is "psychological hedonism"[48]. This hypothesis states that "the only ultimate desires that people have are attaining pleasure and avoiding pain" (Sober, 2013, p. 150). Sober and Wilson (1998) argue that, of all forms of psychological egoism, this is the most difficult to refute and the most likely to have evolved (p. 297). They believe that other egoistic desires seem to be reducible to hedonistic desires (for an objection, see Rottschaefer, 2000). Psychological hedonism became the standard kind of psychological egoism in the literature.

Sober and Wilson (1998) understand that altruistic and egoistic motivations trigger helping behaviors differently because they require different sets of beliefs and instrumental desires in order to motivate behavior (p. 312). If psychological hedonism is true, parents will need an inferential causal chain, with certain beliefs and desires, linking helping offspring with certain self-benefit. For example, if they are helping as a means to obtain pleasure, they depend on the belief that helping leads to pleasure, on an instrumental desire to help, on the presence of actual feelings of pleasure when the desire is satisfied, etc. This inferential causal chain of egoists, with auxiliary beliefs, instrumental desires, and certain feelings, is very complex. By contrast, individuals with altruistic motivation will have a more economic inferential causal chain — all they need to trigger the helping behavior is the belief that the other is in need, and this suffices to trigger helping behavior. Altruistic motivation is cognitively simpler, being more easily activated, less likely to fail, and thus being more likely to produce helping behavior than its egoistic counterpart.

But if the explanation above is right, and altruists produce helping behaviors so easily, then the suggestion that altruistic motivation is adaptive might sound odd. Natural selection

[48] Feinberg (2013) prefers the term "psychological egoistic hedonism".

usually selects behaviors that increase the agent's *own* fitness, not behaviors that increase the fitness of others. Organisms that help others regardless of the benefit they receive are likely to be exploited by free-riders (see Trivers, 1971; West et al., 2007). Thus, if egoistic desires allow us to help others conditionally, keeping track of our best interest, then they seem more at home with natural selection than their altruistic alternative. Egoistic motivation is the motivational mechanism *prima facie* more likely to be selected.

Although Sober and Wilson (1998) agree that egoistic motivation is the evolutionarily optimal strategy in most cases, they argue that altruistic motivation offers an advantage at least in one case: it is more adaptive as a motivation to provide *parental care*[49] (p. 301). They claim that a motivational mechanism that increases the likelihood of helping offspring will have a substantial positive selective pressure. This is especially the case in the human species. As Schulz (2018) comments, "human infants are extremely helpless and completely dependent on adult help" (p. 198). So, although altruistic motivation is not always adaptive, it might be adaptive at least as a proximate mechanism for providing parental care in humans. Thus, the evolutionary argument for psychological altruism is based on the selection of motivation for parental care.

In this introductory section, I have provided an overview of the evolutionary approach to altruistic and egoistic motivation. The next section presents Sober and Wilson's (1998) argument in more detail. There, I also address a second evolutionary argument for psychological altruism, originally proposed by Schulz (2016). After presenting both evolutionary arguments, I will raise a series of objections to them. I will focus mainly on Sober and Wilson's argument, due to its

[49] Batson (2011) also believes that the basis of altruistic motivation should be found in parental care (p. 4).

influence on the literature. *Section 5.3* discusses the costs of altruistic motivation for parents, arguing that it is not clear that these costs cannot outweigh the benefits of altruistic motivation. *Section 5.4* discusses the costs of having different ultimate desires, which are likely to impose another selective pressure against altruistic motivation. Finally, *Section 5.5* discusses how egoistic motivation can avoid some of the criticism raised against it, being as reliable as altruistic motivation. The conclusion I will reach at the end of this chapter is similar to the conclusion reached in the last chapter, namely, that the arguments do not offer a substantial case for psychological altruism.

Before moving to the next section, it is important to make it clear that the altruistic motivation for parental care is not a case of *evolutionary altruism*. The argument discussed here (that altruistic motivation is adaptive as a proximate mechanism) claims that altruistic motivation is selected for increasing the *parent's own direct fitness*. So, the evolutionary case for psychological altruism does not depend on arguments for the selection of evolutionary altruism. As I explained in Chapter 2, evolutionary altruism occurs when a behavior/phenotype benefits another individual and is costly to the performer, where cost and benefit are measured in terms of lifetime fitness (West et al., 2007). Parental care does not qualify as evolutionary altruism for one's fitness is not measured on the basis of the total number of offspring being *born*, but on the basis of the number of offspring *surviving to adulthood*[50] (West et al., 2007, p. 418). As

[50] The scope of one's direct fitness could be extended even further. The standard account of fitness considers a one-generation time frame, that is, it considers only one's offspring. However, as Sober (2001) discusses, this is problematic for some cases. For example, for two individuals that produce the same number of offspring, but with a different sex-ratio (e.g., one produces more females than the other), looking at only one generation would not show the fitness impacts of having such a biased sex-ratio: the advantage of having a certain sex-ratio will only be evident in the second generation (Sober, 2001, p. 30). This might suggest that a proper account of direct fitness would include more than one generation. Authors like Schulz (2018) adopt the view that one's fitness should include not only one's offspring, but also one's grand-offspring (p. 197).

Hamilton (1964a) explains, "[s]acrifices involved in parental care are a possibility implicit in any model in which the definition of fitness is based, as it should be, on the number of adult offspring" (p. 1). So, helping one's own offspring increases one's own direct fitness, and is not a case of evolutionary altruism. Thus, the discussion of the selection of altruistic motivation should not be confused with the discussion of the selection of evolutionary altruism.

5.2 The Evolutionary Case for Psychological Altruism

The evolutionary argument proposed by Sober and Wilson (1998) claims that altruistic motivation is adaptive because it is more *reliable* than egoistic motivation. In their argument, the authors defend two different theses. The first one says that altruistic motivation is more reliable than egoistic motivation as a proximate mechanism to provide parental care. In this first thesis, they compare altruistic and egoistic motivations, asking which one is a better proximate mechanism. The second thesis is slightly different, claiming that *having both* altruistic *and* egoistic ultimate desires is more reliable than having only egoistic motivation. Both theses lead to the conclusion that having ultimate altruistic desires in one's repertoire is more reliable than having exclusively ultimate egoistic desires.

In order to discuss the complex interaction of selective forces acting on the selection of altruistic motivation, Sober and Wilson (1998) discuss, as an example, the selection of proximate mechanisms to avoid oxygen in marine bacteria such as *Aquaspirillum Magnetotacticum* (p. 304). Since oxygen is toxic for these organisms, decreasing the oxygen concentration in their environment is correlated with an increase in their fitness (Blakemore & Frankel, 1981). Sober and Wilson compare two different proximate mechanisms that marine bacteria can use to avoid

oxygen. The first mechanism (*aerotaxis*) is a device that detects oxygen and triggers the behavior of moving away from it. The second mechanism (*magnetotaxis*) detects the earth's magnetic field, using it as a guide to swim to the bottom of the ocean, where there is less oxygen[51].

The relevant difference between the two mechanisms above, for the purposes of the argument, is that *aerotaxis* is a *direct mechanism* (D) and *magnetotaxis* is an *indirect mechanism* (I). The former detects the fitness-enhancing phenomenon (oxygen) directly, while the latter detects something that is correlated with it. This difference between direct and indirect mechanisms is what Sober and Wilson (1998) will use later to discuss altruistic motivation. But before talking about motivation, consider the implications of the direct (D) and indirect (I) mechanisms. Sober and Wilson (1998) propose two different principles that make predictions about them. First, the *D/I Asymmetry* principle:

> D will be a more reliable guide than I concerning which behaviors are fitness-enhancing, if D detects oxygen at least as well as I detects elevation and oxygen and elevation are less than perfectly correlated. (Sober & Wilson, 1998, p. 306)

Since oxygen and depth are less than perfectly correlated, the direct mechanism is likely to be more reliable than the indirect mechanism[52]. In other words, the direct mechanism is likely to produce the behavior of avoiding oxygen in a broader range of scenarios, being more reliable than its alternative. The second principle is the *Two Are Better Than One*:

---

[51] This second mechanism is possible due to an organelle called a *magnetosome*, which consists of intracellular structures containing magnetic minerals that work as a compass, allowing the organism to be oriented by magnetic fields (Frankel & Bazylinski, 2002).

[52] A simpler example of an indirect mechanism analogous to the magnetosome is a mechanism to avoid light. Considering that there is a correlation between light and oxygen, since both are more abundant near the surface, bacteria can use light as a reference to move away from oxygen (Lefèvre & Bazylinski, 2013; see also Frankel & Blakemore, 1989).

> The two devices D and I acting together will be a more reliable guide concerning which behaviors are fitness-enhancing than either D or I acting alone, if each device is positively, though imperfectly, correlated with oxygen level, and if the two devices operate with a reasonable degree of independence of each other. (Sober & Wilson, 1998, p. 307)

The idea is that, if one mechanism fails, the other can still trigger the behavior — assuming that the two mechanisms are independent of each other. Thus, it is more reliable to have two mechanisms than to have only one.

How can the discussion above illuminate our understanding of the selection of altruistic motivation? There is a parallel between the mechanisms responsible for avoiding oxygen in marine bacteria and the motivational mechanisms responsible to provide parental care. Sober and Wilson (1998) claim that *altruistic motivation is a direct mechanism* for providing parental care, while *egoistic motivation is an indirect mechanism* (p. 317). This is so because altruists will trigger help once they believe offspring are in need. Like the oxygen detector in marine bacteria, altruistic motivation responds directly to the phenomenon that increases the organism's fitness. By contrast, individuals with egoistic motivation have to go through a longer inferential process, where they depend on a set of beliefs, instrumental desires, and feelings in order to link helping behavior to their egoistic ultimate goals. These egoists rely on the correlation between parental care and some personal benefit, which makes egoistic motivation an *indirect* mechanism for parental care. Like the detectors of the earth's magnetic field in marine bacteria, egoistic motivation depends on a more complex set of correlations and intermediary steps in order to produce the fitness-enhancing behavior.

Considering altruistic motivation as a direct proximate mechanism and egoistic motivation as an indirect proximate mechanism, we can apply the two principles discussed earlier. By applying the *D/I Asymmetry* principle, it follows that altruistic motivation is more

reliable than egoistic motivation. And by applying the *Two Are Better Than One* principle, it follows that having both altruistic and egoistic motivations in one's repertoire is more reliable than having only egoistic motivation.

The principles above suggest that altruistic motivation is more *reliable* than egoistic motivation as a proximate mechanism to provide parental care. However, the fact that a mechanism is reliable does not entail that it is *adaptive*. There are many evolutionary pressures acting on the selection of any trait, not only its reliability. Sober and Wilson (1998) state that, if we want to compare proximate mechanisms in order to establish which one is more adaptive, we need to consider their *availability*, *reliability*, and *energetic efficiency* (p. 305). Each of these factors represents an independent evolutionary pressure. It is possible, for example, that one proximate mechanism is more reliable but fails to be selected for its high energetic cost or for not being available in the gene pool. An evolutionary argument has to take into account all three factors before concluding which mechanism is more adaptive.

Sober and Wilson (1998) argue that the superior reliability of altruistic motivation, in this case, is sufficient to make it more adaptive than egoistic motivation. They believe that altruistic and egoistic motivations are both available and have virtually the same energetic cost (Sober & Wilson, 1998, p. 321). Regarding availability, Sober and Wilson claim that desires to help others are already available for egoists as instrumental desires. So, all that altruistic motivation requires is for these existent desires to become ultimate[53]. Regarding energetic efficiency, they argue that

[53] The assumption made by the authors about the availability of altruistic desires is very important. Without this assumption, one could ask: how can we defend that altruistic desires exist *because* they are selected if the selection process itself already assumes the existence of these desires? The strategy used by Sober and Wilson (1998) to avoid this problem is to assume that desires to benefit others were already available for egoists, since these desires to help were used as means to egoistic ends. Without this assumption, the evolutionary argument for psychological altruism could be accused of committing a *petitio principii*.

it is unlikely that having more beliefs and desires will demand significant amounts of energy. Thus, energetic efficiency is not a factor in selecting one of the two motivations. They conclude that, since the availability and the energetic efficiency can be dismissed in this case, the fact that altruistic motivation is more reliable is sufficient to justify the claim that it is more adaptive than egoistic motivation.

Until this point, this section was dedicated to presenting a summary of the evolutionary argument proposed by Sober and Wilson (1998). Now, I address the second evolutionary argument, originally proposed by Schulz (2016, 2018) and further developed by Piccinni and Schulz (2018, 2019). This argument also defends that altruistic motivation is more adaptive than egoistic motivation. But it does so in a different way. Sober and Wilson (1998) claim that altruistic motivation is adaptive for being more *reliable*. By contrast, in the argument proposed by Schulz (2016), altruistic motivation is considered to be more adaptive because it is more *cognitively efficient* than egoistic motivation.

Schulz (2016) argues that egoistic motivation demands more of the individual's cognitive resources than altruistic motivation. This is so because the complex inferential mechanism of egoistic motivation is more demanding than that of altruistic motivation. As I said, egoistic motivation will demand a set of auxiliary beliefs to, for example, link helping behavior to an egoistic reward. By contrast, "instead of reasoning about whether to help their offspring, [parents with altruistic motivation] let their perception or belief that their offspring is in need of help simply trigger their ultimate desire to help that offspring" (Schulz, 2016, p. 20). Altruistic motivation triggers help more quickly and uses less cognitive energy, which confers an evolutionary advantage for altruistic motivation in comparison to egoistic motivation. If Sober

and Wilson (1998) are right, and the other evolutionary pressures can be dismissed, Schulz's argument leads to the conclusion that altruistic motivation is adaptive.

It is important to note that Schulz is *not* claiming that altruistic motivation is more *energetically* efficient. As I said, in Sober and Wilson's (1998) argument, it is assumed that altruistic and egoistic motivation will have virtually the same energetic costs. But the point that Schulz (2016) is making is that egoistic motivation, by relying on more representational inferences is more *cognitively* costly. That is, it demands more cognitive resources, such as *concentration* and *attention* (Schulz, 2016, p. 19). Whether having more beliefs and desires significantly increases the organism's metabolic demands is a separate question (see Lemos, 2004).

Although the argument proposed by Schulz (2016) also focuses on parental care, he claims that altruistic motivation could be adaptive in a broader set of contexts. In animals that form enduring social bonds, a simpler mechanism to provide help might be also advantageous. Seyfarth and Cheney (2012) argue that, in long-lived mammals such as primates, dolphins, and elephants, these bonds increase reproductive success in males and reduce stress, increase infant survival, and increase longevity in females. Altruistic motivation could be an adaptive motivation to help in cases where one has this sort of bond. When these stable bonds are formed, it might be adaptive to avoid the cognitive costs of thinking about whether to help (Piccinini & Schulz, 2018, p. 16).

Despite their differences, Schulz's argument is compatible with Sober and Wilson's argument. These arguments can be jointly considered as parts of an evolutionary case for psychological altruism. However, evidence from evolutionary psychology is far from indisputable — even the most influential discoveries in this area have been heavily criticized (see

Buller, 2005). Considering the limitations of this approach, all that the evolutionary arguments try to do is to present evidence in favor of psychological altruism (Sober & Wilson, 1998, p. 323; Schulz, 2018, p. 194). This evidence is expected to be further supported following the principle of *consilience,* where evidence from distinct areas is combined to support a thesis. The evolutionary argument, thus, although weak by itself, could be part of a substantial case for psychological altruism (Schulz, 2016, p. 18). In the following three sections, I raise some problems for the evolutionary arguments, which might preclude them from providing such relevant evidence for a general case for psychological altruism.

5.3 The Selective Pressures Against Altruistic Motivation

Sober and Wilson's argument for the evolution of altruistic motivation was presented in the book *Unto Others* (1998). This book was divided into two parts. The first part covered the contentious issue of group selection in evolutionary biology, providing an argument for the selection of evolutionary altruism through group selection. The second part presented the evolutionary argument for psychological altruism discussed here. These are two very complex arguments, which are independent of each other. The first part, however, because it presents a controversial case for group selection, received the vast majority of the attention. A consequence of this is that the first part of the book obfuscated the second part. As Wilson and Sober (2002) say, "[i]nterest in the group selection controversy often seems to overshadow the equally important issues surrounding psychological altruism" (p. 723). This is reflected in the relatively superficial criticisms it received in the initial reactions to the book.

First, Harman (2000) raises an objection based on a misunderstanding of Sober and Wilson's argument. He assumes that the authors defend that empirical evidence cannot *in principle* support psychological altruism. But as Sober and Wilson (2000) respond, all they claim is that *current* evidence fails in supporting psychological altruism (p. 265). Batson (2000) only argues that Sober and Wilson's dismissal of his experiments is unjustified. He also does not discuss the argument proposed. Rottschaefer (2000) focuses on secondary issues, such as whether the egoistic hypothesis should be focused on hedonistic states or not. Finally, Jamieson (2002) also shows some fundamental misunderstandings, such as the view that Sober and Wilson deliberately invoke folk psychology as a means to make their case. But the authors are simply following the standard approach to the debate on psychological altruism, which is articulated in terms of beliefs and desires.

In their responses to the commentaries above, Sober and Wilson did not discuss the core of their argument for psychological altruism. Their responses ended up focusing mostly on clarifications about the framework of the psychological altruism debate than the argument itself (Sober & Wilson, 2000; Wilson & Sober, 2002). It took some time for a substantial discussion of the argument to be proposed (e.g., Schulz, 2011; Stich, 2007; Stich et al., 2010).

This section and the next aim to contribute to this discussion by pointing out some limitations of the evolutionary argument. The criticism that I will raise in this section is that there are significant evolutionary pressures pushing parents away from having altruistic motivation as a proximate mechanism for parental care. I will discuss how altruistic motivation is costly and how these costs overshadow the benefits that may follow from altruistic motivation.

As explained in the previous section, Sober and Wilson (1998) assume that parents with altruistic motivation produce help whenever they believe that their offspring are in need, while

egoistic motivation demands a more complex and demanding deliberation process. Parents with altruistic motivation are more likely to provide parental care than parents with egoistic motivation. It is argued that this represents a benefit for parents with altruistic motivation. However, a problem that did not receive enough attention is that the unconditional helping of altruistic parents will also represent a significant cost. I challenge the assumption that the benefit of having altruistic motivation *outweighs* the costs of having such a motivation.

Robert Trivers (1972) coined the term "parental investment" to represent "any investment by the parent in an individual offspring that increases the offspring's chance of surviving (and hence reproductive success) at the cost of the parent's ability to invest in other offspring" (p. 139). There is always at least one optimal distribution of parental investment across offspring. If parents invest *more* than the optimal amount in one offspring, then this extra investment is costly, that is, it reduces their fitness. Parents with multiple offspring need to distribute their resources, which sometimes requires neglecting the needs of some offspring (see Schulz, 2018, p. 199). Variables such as the availability of resources or the age of each offspring are relevant for the parents' decision of whether, when, and to what degree they should help.

Since providing help to offspring can be costly, we should expect a *conflict of interest* between parents and offspring. Although parents have an interest in enhancing the offspring's fitness, there is a conflict between how much care offspring demand and how much care parents are willing to provide. This conflict has made natural selection push offspring to try all sorts of strategies to receive as much attention as possible (Dawkins, 1976/2006). Consider, for example, the offspring crying for help as a signal that triggers parental care. If crying has this effect, it can be used by offspring not only to *communicate* their needs but to *manipulate* their parents into believing that they are in need. The behavior of "crying as if you were very hungry" can be a

good strategy to get extra food, regardless of how hungry you actually are. In this case, crying becomes a *social behavior* in which offspring engages in order to *deceive* parents (Dawkins, 1976/2006, p. 131). The crucial point, here, is that if the parent-offspring relationship is a social interaction between agents with conflicting interests, then parents with altruistic motivation are more likely to be *exploited* by their offspring than parents with egoistic motivation. This is so because the former will be more judicious regarding when to help than the former. This constitutes selective pressure against altruistic motivation.

Another important aspect of parental investment is that it takes into account not only current offspring but also *potential* offspring. The best strategy for parents is not simply to optimally divide resources between current offspring but to optimize their parental investment, which includes promoting the interest of future offspring. Taking into account the interest of future offspring will, sometimes, demand denying care for current offspring. In a situation where food is scarce, a starving mother choosing to eat instead of feeding her offspring is acting in the best interest of her future offspring. Thus, ignoring the requests of one's offspring and behaving egoistically, in some cases, is the best strategy to optimize parental investment. Altruistic motivation, by promoting helping behavior directly, is likely to preclude parents from making the optimal choice in these circumstances.

Finally, we need also to consider the competition *between parents*. As Trivers (1972) explains, it is the relative parental investment of each sex that governs sexual selection (p. 141). In sexually reproductive species, where both parents share an equal genetic interest in the offspring, parents compete against each other to invest less in their offspring than the other parent (Dawkins, 1976/2006). The more one sex invests in the offspring, the more this sex selects the mates, and the less one sex invests in offspring, the more it will have to compete for

the mate. This offers yet another selective pressure against over-providing, which is likely to follow from altruistic motivation.

Sober and Wilson (1998) do not provide a response to the problem raised above. However, a possible response can be found in the work of Schulz (2016, 2018). Schulz (2016) claims that, for parents, "not helping their offspring when it would be adaptive to do so is more adaptively costly than helping them when it would not be adaptive to do so" (p. 6). For Schulz (2016), parental care "is sufficiently often the adaptive response to have the disposition to help their offspring stem from an altruistic motivational structure be adaptive" (p. 6). The idea is that the costs of the maladaptive helping provided by altruistic motivation are unlikely to outweigh its benefits. Schulz (2018) adds that "'overproviding' help to human infant offspring is much less of a worry than not providing help when it is needed" (p. 198).

Although Schulz recognizes the problem, I do not thing his answer is satisfactory. After the issues considered in this section, it is far from obvious that the costs of overproviding can be dismissed so quickly. Why are the few cases where egoistic motivation presumably fails in triggering help so overwhelmingly detrimental to the point of justifying a motivation that is constantly making parents over-provide? The answer Schulz provides needs empirical support. Since neither Schulz nor Sober and Wilson provides such evidence, we should minimally suspend our judgment regarding which motivation is adaptive.

Schulz also proposes a different response to the problem:

> [E]ven if it is adaptive for a certain population of organisms to be altruistically inclined when it comes to their offspring, members of the population need not always decide to help all of their offspring. This is due to the fact that, if an organism has multiple offspring in need of help, the decision how to apportion the available resources to these offspring is non-trivial. Indeed, it is consistent with an organism acting on a conative representation [desire] to help (all of) its offspring for it to at times decide to let some of its offspring die — for it may be that the

> best way to help its offspring is to give all of the available resources to some of its offspring, at the expense of some of its other offspring. (Schulz, 2018, p. 198)

In the passage above, Shulz provides a solution to the problem: parents will not over-provide because they are aiming for the best outcome for the majority of offspring. This could be considered an answer in support of psychological altruism. The problem is that it is not really talking about the same ultimate desire based on which the evolutionary arguments are built. Schulz's (2018) passage talks about a desire to optimize an abstract value, namely, the welfare of all of its offspring[54]. But a desire to maximize the overall welfare of all of one's offspring is something quite different from an ultimate desire to increase the welfare of a particular individual. Let me explain why this new form of altruistic motivation is problematic for the evolutionary argument.

Consider, first, how this account of altruistic motivation as a motivation to optimize the welfare of the majority of offspring is problematic for Schulz (2016, 2018). His argument is that altruistic motivation is more *cognitively efficient* than egoistic motivation. But if altruistic motivation is understood as in the passage above, it elicits a quite complex inferential calculus. The calculation of how to optimally allocate resources to promote the maximization of the welfare of all (present and future) offspring is overwhelmingly more cognitively complex than a straightforward desire to help one's offspring in need. It is not clear whether such a complex calculation would be less demanding from egoistic motivation. In fact, it seems quite plausible that this kind of altruist will demand an even more sophisticated inferential calculus than egoists. So, if altruistic motivation is defined as this highly abstract and demanding optimization of the

[54] Furthermore, following Sober (2013), we can even put into question to what degree this motivation is altruistic, for it is not directed to specific others (see Sober, 2013, p. 150). I will discuss this idea in Chapter 8.

welfare of offspring, it is no longer clear that altruistic motivation will be less cognitively efficient, which was a condition for Schulz's evolutionary argument.

Consider how the motivation to optimize the welfare of the majority of offspring would also undermine Sober and Wilson's (1998) argument. Their evolutionary argument is based on the idea that altruistic motivation is more reliable because it is *simpler* than egoistic motivation. But if altruists are aiming to optimize an abstract value, which considers both existing and non-existing offspring, it is no longer clear why we should regard altruistic motivation as simpler than egoistic motivation. As said above, these parents will have to go through a complex inferential calculus, possibly more complex than that of parents with egoistic motivation. Furthermore, assuming that parents are sub-optimal calculators of offspring's welfare, they are vulnerable to a series of mistakes, such as consistently acting selfishly as a means to produce future offspring. These parents might be worse at providing parental care than parents with egoistic motivation. Thus, both evolutionary arguments would fail if we adopted the account of altruistic motivation as a motivation to optimize the welfare of the majority of offspring. The next section adds more reasons for questioning the evolutionary argument. There, I discuss further selective pressures acting against altruistic motivation.

5.4 The Incommensurability of Altruism

In the previous section, I argued that parents with ultimate desires to help their offspring will often produce maladaptive helping behaviors, which produces selective pressure against altruistic motivation. This section discusses a different problem for altruistic motivation, namely, the difficulties of *adjudicating* when ultimate desires are in conflict (see Berman, 2003, p. 148).

Parents who only desire ultimately to increase their pleasure will have only to decide what instrumental desires lead to more pleasure. Parents who ultimately desire pleasure and ultimately desire to help offspring should have a further way to decide which of these two desires to follow when they conflict. This problem is raised by Sober and Wilson (1998) in the following passage:

> A pluralistic organism [with both egoistic and altruistic ultimate desires] may find itself in situations in which its ultimate desires conflict. If the organism is hungry and its child is too, how is the organism to decide what to do if there is not enough food for both? The pluralist [altruist] needs a mechanism for adjudicating. (p. 323)

Consider, first, how a parent with egoistic motivation would solve the conflict above. Since their ultimate desires are all egoistic, both desires to eat and feed their offspring will be instrumental. More precisely, in the case of psychological hedonism, parents will ultimately desire to increase pleasure and reduce pain. So, parents with this egoistic motivation can compare the two instrumental desires and determine which of them is more likely to promote what is desired ultimately. Whenever parents with egoistic motivation desire incompatible goals, they can rely on their representational decision-making to adjudicate the conflict and decide which one to follow. But in the case of parents with altruistic motivation, things are more complicated.

For a parent with altruistic motivation, the desire to eat and the desire to feed offspring are not reducible to the same ultimate desire. So, parents cannot rely on the same inferential process that egoists do. The desire to eat is instrumental to an ultimate egoistic desire, while helping follows from an ultimate altruistic desire. The two conflicting ultimate desires do not share a common value based on which altruists could *rationally* compare them. So, "the altruist will use a non-reasoning-based process to determine which of its ultimate goals to pursue:

different situations will 'trigger' different ones of her ultimate desires to be the determinant of her actions" (Schulz, 2016, p. 18). Therefore, the altruistic parent has a problem of *incommensurability* between different ultimate desires: whenever ultimate desires conflict, altruists will not have the means to rationally decide which desire to follow.

To sum up, we can conclude that when facing the question of whether to eat or feed their offspring, egoists can rely on the comparison of which of these instrumental desires produce more pleasure. For altruistic parents, however, these two actions are based on distinct ultimate desires, which are not trying to obtain something else: having pleasure and helping offspring are not instrumental. So, as Sober and Wilson (1998) comment, parents with altruistic motivation require a mechanism for adjudicating conflicting ultimate desires. The problem for proponents of psychological altruism is that this imposes extra costs on altruistic parents.

If altruists demand an extra mechanism to adjudicate conflicts between ultimate desires, and this is not required by egoists, then the costs of this new mechanism should also be considered in the evolutionary argument. The extra cost of creating and maintaining such an extra adjudicating mechanism results in another selective pressure *against* altruistic motivation. This undermines an important premise of Sober and Wilson's evolutionary argument, namely, the assumption that altruistic motivation does not represent extra energetic costs.

The assumption that altruistic motivation does not produce extra energetic costs is also put into question by Lemos (2004, 2008). Lemos (2004) claims that ultimate desires might be dependent on brain modules that might differ from that on which instrumental desires depend (p. 513). It is a common assumption in evolutionary psychology that the psychological functions, such as the mechanisms controlling behaviors, are individualized into distinct brain modules, which evolved relatively independently (Buller, 2005, p. 63). These modules would demand

energy, thus compromising the assumption that altruistic motivation does not represent extra energetic costs. Lemos (2004) claims that Sober and Wilson dismiss the energetic costs of altruistic motivation way too quickly: if they require distinct brain modules, then we should expect extra costs[55]. The same costs for a distinct module can also be applied to the adjudicating system that altruistic motivation demands.

After introducing the problem of conflicting ultimate desires, Sober and Wilson (1998) propose a solution. They claim that, although altruistic motivation demands an adjudicating mechanism, egoistic motivation also requires one. The adjudicating mechanism for egoists, however, does not compare different desires, but "different types of pleasure and different types of pain" (Sober & Wilson, 1998, p. 323). The authors claim that, if psychological hedonism is true, individuals will still have to compare different kinds of pleasures and pains, which would require an adjudicating mechanism. Sober and Wilson (1998) claim that, since both altruistic and egoistic motivations require one adjudicating system, this should not be used as a criterion for comparing which one is more costly. I will raise two objections to their response.

The first objection to Sober and Wilson's response is that it is not obvious that egoistic motivation demands an extra adjudicating mechanism. We can postulate that individuals with hedonistic ultimate desires attribute *values* to their instrumental desires, depending on how good these instrumental desires are in promoting their hedonistic goals. These values can be simplified in a single scale, going from "positive" to "negative" values, in which pleasures are positive and pains are negative (Schroeder, 2004, p. 73). With this system, egoists can evaluate their desires

[55] Lemos (2004) also questions the availability of altruistic motivation. He claims that the existence of instrumental desires to help does not mean that these are available as ultimate desires. But, again, if these instrumental desires rely on different brain modules, having instrumental desires to help does mean that ultimate altruistic desires are available.

and make the best choice (for a discussion on having a cardinal measure for desires, see Pollock, 2006). Even if imperfect, this can work relatively well in one's decision-making, and there is no reason to doubt why this is, *in principle*, impossible.

The second objection to Sober and Wilson's claim is that *even if* individuals with egoistic motivation need an extra mechanism to adjudicate between different pains and pleasures, so will individuals with altruistic motivation. Remember that individuals with altruistic motivation also have egoistic motivation, and this will include desires to avoid pain and increase pleasure. So, *if* a mechanism to compare different sorts of pains and pleasures is really needed, it will be equally required for parents with egoistic motivation and parents with altruistic motivation. Therefore, this mechanism does not constitute an extra cost for parents with egoistic motivation when compared to parents with altruistic motivation. The parent with altruistic motivation will require *both* the adjudicating mechanism for hedonistic states *and* the adjudicating mechanism for different ultimate desires[56]. Thus, altruistic motivation ends up being more costly.

In the previous section, I discussed parental investment and how it makes altruistic motivation significantly costly. Since the behavior of helping offspring is not always fitness-enhancing, the motivational mechanism for controlling parental care is unlikely to be selected simply on the basis of how good it is to trigger helping behavior. Under certain conditions, altruistic motivation can cause individuals to provide help when the optimal strategy would be to act egoistically. This is a problem for Sober and Wilson (1998) and Schulz (2016) in their case for psychological altruism. In this section, I added a new cost for parents with altruistic motivation: they need a mechanism to adjudicate when their ultimate desires conflict.

[56] Following Lemos (2004), I assume that the mechanisms required to adjudicate between different pains is *different* from that required to adjudicate between different ultimate desires (p. 516).

Before ending this section, I will address what I believe to be a possible objection to the criticism raised here and in the previous section. One might accuse me of committing a *strawman fallacy*: the idea of parents blindly providing care might be an uncharitable understanding of how altruistic motivation works. Parents with altruistic motivation will certainly have some mechanisms to regulate when providing care and when not doing so. They might regulate helping based on the current emotions they are feeling, such as empathic concern; they might form secondary desires to not help when helping is particularly costly; and so on. To think that the authors addressed here hold this view of parental care is to have a very uncharitable (and wrong) interpretation.

My response to this objection is that, although these different mechanisms to regulate helping behavior are possible, they are *not considered* in the evolutionary comparison between altruistic and egoistic motivations that Sober and Wilson (1998) and Schulz (2016) present. Altruistic motivation is said to be adaptive due to its simplicity: it triggers help whenever parents believe that their offspring is in need. But if it needs a set of auxiliary mechanisms to do its job, then it is no longer obvious that it is that simple. It is certainly possible for one to argue that, considering all these auxiliary mechanisms, altruistic motivation is still more adaptive. But that would be a different argument.

My criticism is not claiming that Sober, Wilson, and Schulz lack knowledge of the basic costs of parental care, but simply that their arguments unjustifiably overlook these costs. In short, they have a dilemma in their hands: either (1) altruistic motivation can avoid maladaptive helping but is no longer simple, thus losing simplicity, which is precisely the trait considered to make it reliable and cognitively efficient, or (2) altruistic motivation is simple and cognitively reliable but is subjected to the criticism raised in sections 5.3 and 5.4, which undermines the

claim that this motivation is adaptive. Both options lead to the same conclusion: the evolutionary arguments fail to make a case for psychological altruism.

5.5 Egoistic Motivation Reconsidered

The evolutionary argument for psychological altruism depicts egoistic motivation as less reliable and less cognitively efficient. The key feature of egoistic motivation that makes it different from altruistic motivation is the long and complex inferential causal chain linking ultimate egoistic desires to helping behaviors. This section analyses this egoistic inferential causal chain. I will discuss three specific problems attributed to the inferential causal chain of egoistic motivation: (1) the need for the presence of feelings; (2) the vulnerability to negative evidence; and (3) the reliance on a long and demanding causal chain linking desires to actions. I will argue that egoistic motivation can plausibly avoid these challenges. The evolutionarily relevant differences between altruistic and egoistic motivation are thus put into question.

The first problem attributed to egoistic motivation is that its inferential causal chain relies on *feelings*. Discussing egoistic motivation, Sober and Wilson (1998) claim that "whenever the [egoistic] organism believes that its children are well off, it tends to experience pleasure; whenever the organism believes that its children are doing badly, it tends to feel pain" (p. 315). They claim that, in order to produce helping behaviors, egoistic motivation needs to have not only beliefs and desires, but also to experience certain feelings. Sober and Wilson (1998) also argue that, in order to produce behavior, these feelings of pleasure or pain have to be particularly *strong*, since they will compete with other motivations. So, considering that these individuals might *fail* in producing these feelings sometimes and that on certain occasions they will fail in

producing these feelings at the correct intensity, it follows that egoists are likely to fail in producing parental care sometimes.

Now, consider how egoistic motivation can avoid the problem above. The egoist depicted by Sober and Wilson (1998) depends on the *occurrence* of certain feelings, and this makes their motivation less reliable than altruistic motivation. However, although the egoistic motivation described by Sober and Wilson is certainly possible, it is not the only option. Psychological hedonism can use different sets of beliefs and instrumental desires that rely on this affective element in a different way. To illustrate how individuals with egoistic motivation do not need an affective component, Stich (2007) proposes an example: parents with egoistic motivation can provide parental care as a means to *avoid* certain feelings. They may believe that not offering parental care would make them feel bad. In this particular arrangement, which Stich (2007) calls *Future Pain Hedonism*, the egoistic parents might never actually feel the feelings of pleasure and pain associated with helping (p. 273). Thus, "it seems (logically) possible to eliminate the pathway that runs via affect" (Stich, 2007, p. 275). The inferential causal chain linking egoistic motivation to helping behavior does not require necessarily the presence of certain feelings. Thus, the occasional failure in producing certain feelings is not as much a challenge for egoists as Sober and Wilson (1998) assume.

The second problem in the inferential causal chain of individuals with egoistic motivation is *the problem of negative evidence*. Egoistic parents have to believe that taking care of offspring produces pleasure or reduces pain. However, if they help offspring and this help fails in providing the feelings they expected, they might *correct* their belief (Sober & Wilson, 1998, p. 314). The possibility of learning that parental care does not produce positive feelings makes

egoistic motivation more vulnerable to failure in providing parental care than altruistic motivation, thus being less reliable than parents with altruistic motivation.

Now, consider how the second problem can be avoided. In Sober and Wilson's (1998) discussion, egoists' beliefs linking parental care to hedonistic states are subjected to revision. But are the beliefs in egoistic motivation necessarily subjected to revision? I will propose two scenarios where this is not the case. The first is that desires can be unconscious. Imagine that a parent helps her offspring moved by an ultimate egoistic desire to feel pleasure. However, she believes to do so out of an ultimate concern for the offspring. That is, she only have access to the instrumental desire to help, while the egoistic desire remains unconscious. In this case, if her experience of helping does not produce pleasure, she might not be able to identify this as a piece of evidence against their motivation to help.

Another scenario that we can consider in order to question the problem of negative evidence is one where an egoist's beliefs are *sub-doxastic states* (Stich, 2007). Sub-doxastic states are basic mental states that function like beliefs but are much more rigid. As Stich (2007) explains, they are belief-like states that are not subjected to change based on contrary evidence (p. 278). As an example of these states, Stich (2007) mentions certain innate grammatical rules. Sub-doxastic states are not so well inferentially integrated with our beliefs (Stich, 1978, p. 506). They are more isolated and more stable. These are not subjected to correction, so they are immune to the problem of negative evidence.

In addition to sub-doxastic states, the notion of *core beliefs*, proposed by Carey and Spelke (1996), is also an example of how beliefs can be resistant to revision. For example, certain core beliefs about physical objects remain stable regardless of contrary evidence. As Carey and Spelke (1996) comment, "core cognitive systems remain not only when they give rise

to beliefs that are true and useful, but also when they do not" (p. 519). Parents' beliefs that taking care of their children is advantageous to themselves can be either a sub-doxastic state or a core belief; in either case, they would not be necessarily subjected to revision.

Sub-doxastic states and core beliefs are mere logical possibilities. But do we have reasons to believe that they are likely to be part of egoistic motivation? I believe that we have good reasons to think so. If Sober and Wilson (1998) are right about the problem of negative evidence, then the belief that taking care of one's offspring is good for oneself would have a strong selective pressure to be as stable as possible. Having a sub-doxastic state of a core belief would be a viable way of obtaining such stability. This would be at least simpler than creating a whole new motivation, as psychological altruism requires[57].

Finally, the last problem attributed to egoistic motivation is that it produces and relies on a long and demanding causal chain of desires and beliefs. But this also can be questioned. Consider *long-standing instrumental desires*, discussed in the previous chapter (Stich et al., 2010, p. 195). These desires can be stable and can be triggered without the need for agents to revisit the whole inferential causal chain linking them to the ultimate desire behind them. If our desires to help our offspring are long-standing desires, then they would be instrumental to egoistic desires, but they would also be, functionally, virtually identical to altruistic motivation. These desires would be easily activated once the agent believes that the offspring is in need. Interestingly, they would also not require much of cognitive resources such as attention and

[57] Schulz (2016) seems to accept that sub-doxastic states offer a legitimate criticism of the evolutionary argument for the superior reliability of altruistic motivation. He states that the existence of sub-doxastic beliefs is widely accepted in the literature, and that "it seems that there is no reason to think that a sub-doxastically motivated egoist either could not exist or that it would not be just as reliable to help others as an altruist would be" (Schulz, 2016, p. 18).

concentration. It is not obvious why these long-standing desires cannot be as reliable and as cognitively efficient as altruistic motivation.

In this section, I responded to some of the problems attributed to egoistic motivation in the evolutionary arguments. Although authors show how egoistic motivation *could* have limitations that would make it less reliable, we do not have good reasons to believe that they do have such limitations, since there are equally possible alternative egoistic explanations. Sober and Wilson (1998) suffer from a similar problem to that which they accuse Batson (1991), namely, they reject only specific forms of psychological egoism but fail in making a case against psychological egoism in general[58]. Egoistic motivation can be structured in ways that avoid the problems attributed to them.

A possible criticism of my argument in this chapter is that the complex combinations of instrumental desires that I have discussed here are far-fetched and implausible. However, I believe that, if psychological egoism is true, we should expect all sorts of complex combinations of instrumental desires. Long-standing instrumental desires to help others, which are stable but ultimately egoistic, are very plausible motivational states we should expect if psychological egoism is true.

My goal in this chapter was not to argue that psychological altruism is false, nor to support psychological egoism. My goal is simply to show how the evolutionary arguments for psychological altruism either fail or provide extremely weak support for the hypothesis. When we take into account the costs of altruistic motivation, as I presented here, we see that the possible benefits that the authors claim to follow from altruistic motivation do not outweigh the

[58] Although this section was focused on Sober and Wilson's views of egoistic motivation, the last problem can also be applied to the argument defended by Schulz (2016).

costs — or at least we do not have enough reasons to believe that they do so. Evolutionary psychology has a challenging object of study, and ultimate desires are a particularly hard case. Perhaps substantive evidence for altruistic motivation would demand much more sophisticated modeling and empirical evidence, going beyond the sort of argument provided by Sober and Wilson (1998) — which can be regarded, as Batson (2000) claims, as armchair evolutionary speculation (p. 209).

This closes Part II of this thesis. In Chapters 4 and 5, I discussed the main cases in favor of psychological altruism in the literature. The conclusion I reached is that both empirical work in social psychology and the evolutionary arguments fail in providing a substantial defense of psychological altruism. Egoistic motivation can have many different inferential causal chains linking ultimate desires to helping behaviors, and these seem to escape both from empirical tests and evolutionary arguments. The conceptual approach to altruistic and egoistic motivation in terms of ultimate desires leads to insurmountable predicaments — being ultimately an unfruitful framework for scientific research. I now move on to Part III, which aims to address the historical development of the idea of altruistic and egoistic motivations. Chapter 6 is dedicated to the history of egoism, while Chapter 7 addresses the history of altruism.

Chapter 6

The History of Egoism

6.1 Egoism and The British Moralists

There is a relatively common historical assumption in the literature on psychological altruism. It is assumed that our contemporary debate, which opposes psychological altruism and psychological egoism, is the continuation of a long-standing debate, dating back at least to modern philosophy. Part III of this thesis looks at the historical development of the notions of egoism and altruism, discussing how the historical accounts are related to the contemporary accounts of altruistic motivation. The present chapter offers a historical investigation of modern views on *egoism*, and the next chapter analyses the history of *altruism*.

The historical assumption above has a *weak* and a *strong* form. In its *weak* form, the historical assumption merely states the proximity between modern views and contemporary views. It indicates how our contemporary debate has its roots in the modern debate, and how some of these authors articulated notions of altruism and egoism that resemble ours in some ways. This is a correct assumption, and will not be disputed. However, in its *strong* form, the historical assumption identifies psychological egoism and psychological altruism as the central views under contention in the modern discussion of egoism. The accounts of human nature proposed by modern authors are then interpreted through the lens of our notions of psychological altruism and psychological egoism. Philosophers such as Hobbes, Mandeville, and Bentham are portrayed as proponents of psychological egoism, while authors such as Shaftesbury, Hutcheson, and Butler are described as arguing that we are capable of holding altruistic ultimate desires.

Many authors discussing psychological altruism assume either the weak or the strong version of the historical assumption (e.g., Clavien & Chapuisat, 2013, p. 127; Garson, 2015, p. 9; Oliner & Oliner, 1988, p. 5; Piccinini & Schulz, 2019, p. 2; Sober & Wilson, 1998, p. 1-2; Stich, 2016, p. 3; Stich et al., 2010, p. 147). It is not always easy to tell whether they adopt the weak or the strong versions, but, at least in some cases, there seem to be good reasons to believe that they adopt the strong version. For example, in the first part of Batson's *The Altruism Question* (1991), he presents the history of modern philosophy as dominated by psychological egoism (p. 22). He claims that, after the Renaissance, the view that humans can only ultimately desire their own welfare is widespread. Hobbes, Mandeville, and Bentham, among others, are presented as proponents of psychological egoism (Batson, 1991, p. 22-25). Batson (1991) holds that "Hobbes and his followers" assume that "human nature is exclusively self-interested", thus holding the strong version of the historical assumption (p. 27).

In this chapter, I criticize the *strong* form of the historical assumption, discussing specifically the conceptions of egoism in some modern philosophers. I will argue that modern authors who are identified as holding an egoistic view are mainly concerned with different forms of egoism, so we should not equate their egoistic account of human nature with psychological egoism. I will also argue that, even though sometimes these authors indicate that they hold a view very close to psychological egoism, psychological egoism does not play a central role in their arguments. Finally, I will defend the claim that we have evidence that these authors accepted the existence of what we now would call altruistic ultimate desires. If this criticism is correct, then the contemporary debate opposing psychological egoism and psychological altruism cannot be justified as being the continuation of the modern debate, for they differ significantly. If the historical legitimacy of the contemporary debate is undermined, then we can

more easily ask the question of whether this is the way we should be thinking about altruism and egoism, which I will discuss in the last part of this thesis.

Each of the next three sections is dedicated to one influential modern philosopher who famously defended what is considered an egoistic account of human nature, namely, Thomas Hobbes, Bernard Mandeville, and Jeremy Bentham. I will discuss the views of these philosophers, showing how psychological egoism is either absent from their work or, at best, plays a secondary role in them. The last section is dedicated to exploring different ways of thinking about egoism, showing how egoism can be conceived in different ways and how psychological egoism is far from being the most plausible or parsimonious version of egoism. These alternative accounts of egoism can better describe the views of modern philosophers.

Before moving on to the next section, let me discuss the philosophical background of the modern debate on egoism and explain why this period is of particular interest. The philosophers discussed in this chapter belong to the tradition of British Moralism. However, we can find the roots of the contemporary debate not only in modern philosophy, but also in Epicureanism (Maurer, 2019, p. 16), Aristotle (Kahn, 1981), and ancient Chinese philosophy (Dubs, 1951). However, there are reasons for focusing particularly on modern philosophers when it comes to the debate on egoism. Much of the work of these modern philosophers was dedicated to proposing a solution to the problem of conflicting self-interests, which was, in a sense, a new problem in philosophy (Darwall, 1995, p. 4). From Aristotle to Aquinas, the pursuit of one's self-interest — one's own *good* — did not entail conflict with other individuals pursuing their own self-interest (Darwall, 1995, p. 5). If one pursued one's own *true good*, there would be no conflict with other individuals pursuing their own true good (MacIntyre, 1967, p. 463). This is so because these different self-interests were part of the same metaphysically cohesive whole:

"[a]ny deep conflict between their goods is thus ruled out — harmony is guaranteed by perfectionist/teleological metaphysics" (Darwall, 1995, p. 5).

In the classical Aristotelian-Thomistic theory, if each one followed his or her own true good, everyone would be better off. But things changed in early modern philosophy. As the classical Aristotelian-Thomistic theory gradually lost its strength, philosophers faced the challenge of finding the basis of civilized society and morality without relying on a metaphysical guarantee of harmony between self-interests. When human nature was presented naked of the metaphysical dress that once covered its true colors, some modern philosophers thought that we would not find the basis of morality in such a bestial human nature. So, alternatively, realizing the strength of egoism in human nature, these philosophers argued that the ultimate roots of civilized society and morality are not the kindness of our hearts, but our selfish instincts. The next section addresses the main icon of this tradition.

6.2 Hobbes and The War of All Against All

The first author analyzed here will be Thomas Hobbes (1588–1679), who Darwall (1995) views as the most original and influential of the early British moralists (p. 53). One of Hobbes's (1642/1983) most influential ideas is that the state of nature is a *war of all against all*, where life is "fierce, short-lived, poor, nasty, and destroy'd of all that Pleasure, and Beauty of life" (p. 49). This view has been taken as a suggestion that, for Hobbes, humans are antisocial creatures whose civility and morality are merely a thin veneer covering a selfish nature (de Waal et al., 2006, p. xi; Wynn et al., 2018, p. 3). In this view, the Hobbesian account of human nature is "essentially individual, nonsocial, competitive, and aggressive" (MacIntyre, 1967, p. 463). A popular

epitome of this interpretation is the ancient Roman proverb "*homo homini lupus*" (a man is a wolf to another man), mentioned by Hobbes (1642/1983, p. 24).

Psychological egoism is often taken as a corollary of the Hobbesian harsh account of human nature (Kavka, 1986, p. 50). "[I]t is relatively common… to present Hobbes's moral psychology as the seventeenth-century paradigm of psychological egoism" (Maurer, 2019, p. 17). Many authors, such as Batson (1991), believe that, in Hobbes's view, "[o]ur ultimate goal or desire is forever and always to benefit ourselves" (p. 23). In fact, we can find in Hobbes's work some passages that seem to endorse psychological egoism. Hobbes (1651/1998) says that "of all voluntary acts, the object is to every man his own good" (p. 100). In *De Cive* (1642/1983), Hobbes claims that "[w]e doe not therefore by nature seek Society for its own sake, but that we may receive some Honour or Profit from it; these we desire Primarily, that Secondarily" (p. 42). In his *Elements of Law, Natural and Politic* (1640/1928), Hobbes defines pity, which is usually considered to be an other-directed passion, as the "imagination or fiction of future calamity to ourselves, proceeding from the sense of another man's present calamity" (pp. 30-31). These passages seem to suggest psychological egoism.

In this section, I will discuss some aspects of Hobbes's philosophy that can shed new light on his account of egoism. I start by presenting the Hobbesian argument for the state of nature as a state of war of all against all, showing how this argument does not entail nor assume psychological egoism, much less an inherently antisocial human nature. The view I defend in this section is not new, but something already accepted by Hobbes scholars (see Lloyd & Sreedhar, 2022). After this, following Gert (1967) and McNeilly (1966), I show how Hobbes, in his late work, explicitly accepts the existence of what we would call today altruistic ultimate desires.

Consider now a summary of Hobbes's argument for why the state of nature is a war of all against all[59].

In *De Cive* (1642/1983) and later in *Leviathan* (1651/1998), Hobbes assumes that humans are concerned *mainly* with their own self-interest. More precisely, humans are concerned chiefly with their *self-preservation* (Hobbes, 1651/1998, p. 111). In the state of nature, where there is no civil society, no laws restrict our behaviors, which makes us entitled to pursue whatever seems better for our preservation (Hobbes, 1642/1983, p. 32; 1651/1998, p. 86). Hobbes (1642/1983) explains that we usually prefer peace over a state of aggression (p. 50). However, desiring the same non-shareable objects invariably puts us in conflict with others (Hobbes, 1651/1998, p. 83).

This scenario becomes worse when we consider how fragile humans are: even the weakest of us have enough strength to kill the strongest (Hobbes, 1651/1998, p. 82; 1642/1983, p. 45). On top of that, we cannot predict *who* is an aggressive individual. Even though Hobbes (1642/1983) thinks that not all humans are wicked, he sees it as reasonable to expect that in any large group *some* individuals will be willing to use aggression to obtain what they want (p. 33). Thus, considering the scenario described here, Hobbes asserts that the best strategy to assure one's safety is *anticipating aggression* (Hobbes, 1651/1998, p. 106; 1642/1983, p. 33). Finally, since every human in the state of nature is likely to adopt this strategy, aggression will be widespread. Hence, the state of nature is a state of *war of all against all*.

[59] This argument is at the base of Hobbes's political philosophy. He argues that the state of nature persists until there is a common power — a *sovereign* — capable of keeping all individuals in awe by threatening them with punishments in case they break the laws (1651/1998, p. 84). Considering our desire for self-preservation, reason entails that one should concede part of one's freedom to a sovereign in exchange for the establishment of the minimal conditions for a peaceful life (Hobbes, 1651/1998, p. 111).

As the summary above shows, the state of war of all against all does not follow from an inherently antisocial nature. What Hobbes argues is that being aggressive is the *rational* thing to do in the uncertain context of the state of nature (McNeilly, 1966, p. 206). "The bestial ethos of the state of nature shows not that humans are really brutish, but that in some circumstances, there is no reasonable alternative to behaving like a beast" (Newey, 2008, p. 55). As Hobbes (1651/1998) explains, when we lock our doors, we are not assuming that everyone is prone to steal from us — we just assume that there are a *few* individuals that might do so (p. 84). In the state of nature, we also do not assume that everyone is wicked and would like to harm us. But since preemptive aggression is the rational strategy to adopt when others might be aggressive and wicked, then rational individuals are likely to behave aggressively. But this is not ascribing a particular feature to human nature in general.

To make sense of why preemptive aggression is the best strategy in the state of nature, some authors have used terminology from game theory. It has been common to interpret Hobbes's state of nature in terms of a *prisoner's dilemma*[60] (Alexandra, 1992). In this game, played by two individuals, one can either cooperate or defect. If both cooperate, both obtain benefits. However, if one cooperates and the other defects, the one that defects obtain a better reward than if both cooperated, while the cooperator suffers a heavy loss. In a prisoner's dilemma, the best outcome is to defect when the other cooperates. Crucially, if the other defects and the agent cooperates, the agent is worse off than if she cooperated. Considering this explanation, we can say that the defecting strategy *strictly dominates* that of cooperating. That is, independently of what the other player decides to do, "one is better off defecting" (Skyrms,

[60] Alexandra (1992) argues that the proper interpretation of the Hobbesian state of nature is an *assurance game*. The difference is that, in the assurance game, the best outcome is the mutual cooperation rather than the "I defect, you cooperate" outcome.

2014, p. 47). In the context of the Hobbesian state of nature, preemptive attacking is defecting, and not attacking is cooperating. Thus, if defecting is the dominant strategy, even otherwise peaceful individuals are likely to act aggressively in the state of nature[61].

The argument summarized here has some problems. Hobbes is often criticized for neglecting other forms of achieving the minimal coordination necessary to assure cooperation and depart from the conflicting state of nature (Lloyd & Sreedhar, 2022). His overly "pessimistic" view rules out alternatives that could lead to the restriction of conflict[62]. As Kavka (1983) points out, Hobbes emphasizes the benefits of preemptive violence, but does not consider in depth the disadvantages of such a strategy (p. 298).

Regardless of the possible flaws in his argument, we can see that Hobbes's argument does not presume that humans are inherently antisocial. The aggressive behaviors follow from one's desire for self-preservation in an environment where aggression is likely. More importantly, psychological egoism is also not necessary. The view that psychological egoism is the foundation of Hobbes's moral philosophy is widely rejected by Hobbes scholars today (see Lloyd & Sreedhar, 2022).

But beyond the view that psychological egoism is just a secondary aspect of Hobbesian philosophy, some authors go further and claim that the interpretation of Hobbes as endorsing psychological egoism is not warranted. A tradition of scholars rejecting the view that Hobbes

[61] Notice that this does not mean that Hobbes thinks this to be the *only* cause of aggression. He accepts that there are plenty of wicked individuals who are likely to attack others out of the pursuit of glory and other selfish motives, but he assumes that these are a minority (Hobbes, 1642/1983, p. 33; 1651/1998, p. 83).

[62] As an example of how one could achieve this minimal coordination, we could mention Frank's (1988) theory of emotions. This theory says that our emotions have the social function of precluding us from performing certain behaviors without our rational command, in a way that is predictable by others. This could generate the basis for coordination.

supported psychological egoism started with Gert (1967) and McNeilly (1966). These authors claim that "though Hobbes's political theory requires that all men be concerned with their own self-interest, especially their own preservation, it does not require that they cannot be concerned with anything else" (Gert, 1967, p. 512).

Although Hobbes does not talk much about non-egoistic desires, he accepts that they are part of our repertoire of desires. In *Leviathan* (1651/1998), Hobbes presents a list of desires, one of which is the "[d]esire of good to another, BENEVOLENCE, GOOD WILL, CHARITY" (p. 37). He does not present this as an instrumental desire. As Ewin (1991) comments, in Hobbes, individuals have multiple ends, and "not all of those ends will be selfish or directed toward the purely private interests of the agent" (p. 117). Since psychological egoism claims that *every* ultimate desire is egoistic, Hobbes's acceptance of desires for the good of another is inconsistent with the view that he holds such a view. As Gert (1967) points out, if Hobbes accepted at least *one* ultimate desire for the good of others, then we cannot ascribe psychological egoism to him (p. 507; see also McNeilly, 1966, p. 203).

When we see Hobbes's emphasis on egoistic motivations, we have to keep in mind that he was a *political* philosopher (see Kavka, 1986, p. 31). When Hobbes discusses how we can leave the state of nature, the motivations that matter are the ones that are powerful enough to move us out of such a state. The *causally relevant* motivations are the egoistic ones. The desire for self-preservation is at the root of the conflicting state of nature and also the basis for the absolute power of the sovereign. Benevolent motivations, by contrast, are seen by Hobbes as limited and incapable of restraining the aggressiveness of the state of nature (Gert, 1967, p. 513). For Hobbes, it is impossible to base civil society on our benevolent passions, hence why Hobbes neglects them in his discussion. However, this approach does not entail that those non-egoistic

motivations *do not exist*. As Hampton (1995) comments, "although the desire for self-preservation is the strongest, Hobbes certainly does not argue that our other desires are derived from it" (p. 17).

As commentators have noted, there are significant *inconsistencies* in Hobbes. These inconsistencies occur between his different books and, sometimes, between claims in the same work — sometimes between claims on the same page (Gert, 1967, p. 507). McNeilly (1966, p. 196) goes as far as to claim that "it is a hopeless project to try to reconstruct some single doctrine of Hobbesian psychology". Nevertheless, McNeilly (1966) proposes that we can read these inconsistencies as signs of progress in Hobbes's work. McNeilly (1966) claims that we can explain most of the passages in which Hobbes seems to endorse psychological egoism as a trait of his early work; a trait that was abandoned in his late more mature work[63].

As evidence for the claim that Hobbes abandoned his early conception of humans as having only egoistic motivations, we can consider the preface for the English edition of *De Cive* (Hobbes, 1642/1983). There, Hobbes (1642/1983) says that "though the wicked were fewer then [sic.] the righteous, yet because we cannot distinguish them, there is a necessity of suspecting, heeding, anticipating" (p. 33). That is, the state of war of all against all is not caused by an inherent antisocial human nature, but by the state of uncertainty characteristic of such a state. McNeilly (1966) claims that Hobbes stated this idea clearly in the preface of the English version

[63] Gert (1967) argues that one of the reasons for the historical association between Hobbes and psychological egoism is the influential criticism proposed by Bishop Butler (1692–1752). In his *Fifteen Sermons* (1726/2006), Butler argues that some of our desires are directed towards external things. He claims that Hobbes holds the view that the only possible desires one might have are self-directed (Butler, 1726/2006, p. 47). But this does not represent Hobbes properly. As McNeilly (1966) comments, "the only thing that is wrong with Butler's famous refutation of Hobbes is that Hobbes thought of it first" (p. 201). The target of Butler is not a proper representation of Hobbes's ideas, but a simplified version of it. In the next chapter, I will discuss Butler's argument.

of *De Cive* as a response to the accusations raised against him of putting forward a view of human nature as inherently antisocial (p. 206). As Hobbes progressed in his work, he gradually abandoned the idea that human motivation is only egoistic, as he suggested mainly in his early work. Kavka (1986) claims that, by the time *Leviathan* was written, in 1651, the radical egoism that seems to suggest psychological egoism was already abandoned. In *Leviathan*, "no significant conclusions are derived from, or presuppose [psychological egoism]" (Kavka, 1986, p. 46).

I argued in this section that Hobbes's philosophy does not depend on (nor imply) psychological egoism. Reading his egoistic account of human nature in terms of psychological egoism impoverishes his actual views, distracting us from the points that are actually important in his work. His arguments rely on more plausible assumptions than psychological egoism and become stronger without assuming such a hypothesis. The next section shows how the same can be said of Mandeville's philosophy.

6.3 Vice and Selfishness in Mandeville

Bernard Mandeville (1670–1733), a Dutch medical doctor who migrated to England, is another author often identified as a proponent of psychological egoism (Batson, 1991, p. 24). The association between Mandeville and psychological egoism is not surprising: Mandeville depicts humans as deeply egoistic creatures, devoid of true virtue. He presents such a view through a fiercely provocative style:

> There is no Merit in saving an innocent Babe ready to drop into the Fire: The Action is neither good nor bad, and what Benefit soever the Infant received, we only obliged our selves; for to have seen it fall, and not strove to hinder it, would have caused a Pain, which Self-preservation compell'd us to prevent.
> (Mandeville, 1714/1988, Vol. 1, pp. 42-43)

In the 18th century, Mandeville's ideas inspired much discussion, mostly by authors opposing his views (Maurer, 2019, p. 80). No doubt with some hyperbole, Sheridan (2007) says that it "is difficult to find a text in the period that does not have something to say in response to Mandeville's egoistic account of human morality" (p. 377). In this section, I will discuss how Mandeville's egoism cannot be reduced to psychological egoism. Mandeville is concerned with what sort of motivation moves us to behave virtuously, not so much about whether altruistic motivation exists or not. Similar to what I argued in the previous section regarding Hobbes, I provide evidence that Mandeville accepted the existence of what we would now consider ultimate altruistic desires.

In his most influential work, *The Fable of the Bees or Private Vices, Publick Benefits* (1714/1988), Mandeville claims that, in the state of nature, individuals are *mainly* concerned with their own appetites, valuing these appetites above all else. He claims that moralists subverted these values, labeling some behaviors that would be weaknesses in nature as morally noble[64] (Mandeville, 1714/1988, Vol. 1, p. 27). But for Mandeville, even though humans often behave contrarily to their more selfish impulses, they learn to do so "motivated by desires for honor and admiration" (Welchman, 2007, p. 59). Civilized individuals deny some instincts but are rewarded with the applause of society. So, individuals considered virtuous in society only change one selfish desire for another. But the virtuous behavior produced in these circumstances is not really virtuous.

[64] Mandeville's genealogy of morality is similar to that proposed later by Nietzsche in his *On the Genealogy of Morality* (1887/1998).

As Maurer (2019) explains, Mandeville distinguishes "true virtue" from "social virtue" (p. 70). Social virtues are the behaviors we observe in society, which are widespread. But true virtue, for Mandeville, is ascetic virtue, which he considers to be either rare or non-existent. True virtue is obtained through the denial of our passions. Mandeville (1714/1988, Vol. 1) describes *virtue* as "every Performance, by which Man, contrary to the impulse of Nature, should endeavour the Benefit of others, or the Conquest of his own Passions out of a Rational Ambition of being good" (p. 34). *Vice*, by contrast, is "every thing, which, without Regard to the Publick, Man should commit to gratify any of his Appetites" (Mandeville, 1714/1988, Vol. 1, p. 34). As Maurer (2019) comments, this is considered to be an "'ascetic' and 'rigorist' conception of morality (that is, a conception based on self-denial in the strong sense of the frustration of our passions)" (p. 70). What we see in society, therefore, is not true virtue, but vice disguised as a virtue.

The first point we need to distinguish is the opposition between altruistic and egoistic motivation and the opposition between passion and self-denial. Considering that Mandeville sees true virtue not as following from a particular passion but from a rational conquest of one's passions, even if altruistic motivation exists, it cannot not be the basis of true virtue. The question of whether one can achieve self-denial is independent of whether one can desire to increase the welfare of others. This is the first point to keep in mind in order to not misinterpret Mandeville's views.

The next issue I will address is how Mandeville was skeptical that our kind motivations to help others could be the base of civil society. The true foundations of society, he claims, are distant from true virtue and kindness. This idea is presented in an allegorical poem called "*Grumbling Hive, or Knaves Turn'd Honest*" (Mandeville, 1714/1988, Vol. 1, pp 1-24). In this

poem, we find the story of a beehive inhabited by bees concerned with their own self-interest and incapable of true virtue. At some point, unhappy with such widespread viciousness, the bees ask the gods to eliminate all vice from their society: "Good Gods, Had we but Honesty!" (Mandeville, 1714/1988, Vol. 1, p. 12), asked the bees. When Jove attends to their wish, however, the bees see all that they value as good vanishing as well. Together with vice, the motor underlying the hive's prosperity was gone.

> *Fraud, Luxury and Pride must live,*
> *While we the Benefits receive:*
> *Hunger's a dreadful Plague, no doubt,*
> *Yet who digests or thrives without?* (Mandeville, 1714/1988, Vol. 1, p. 23)

The moral of the poem is that society is ultimately moved by vice, and, without it, it would collapse[65] — "the world is providentially ordered so that we can live in a peaceful and flourishing society of hypocrites and flatterers" (Maurer, 2019, p. 210). Mandeville argues that people *behave* virtuously because they believe that they will be rewarded for doing so. People were led into believing that "it was more beneficial for every Body to conquer than indulge his Appetites, and much better to mind the Publick than what seem'd his private Interest" (Mandeville, 1714/1988, Vol. 1, p. 28).

To understand the work of Mandeville, we have to consider another author, who was relentlessly criticized by him, namely, the third Earl of Shaftesbury (1671–1713). Shaftesbury (1711/2001) argued against what he considered to be an overly egoistic view of human nature, which became widespread after Hobbes. Shaftesbury's work became one of the first and most

[65] Mandeville is sometimes interpreted as an economic theorist (Maurer, 2019, p. 81). We can see in his work a rudimentary version of Smith's idea of the Invisible Hand: a society of individuals following their own self-interest can produce a harmonious and prosperous society (see Mandeville, 1714/1988, Vol. 1, pp. 80-81).

influential modern arguments for a more sociable human nature. He argued that humans have strong "social affections" based on which they act virtuously, looking for the good of others. But Mandeville was not convinced by it, and strongly criticized Shaftesbury's views. Mandeville (1714/1988, Vol. 1) comments that Shaftesbury "Fancies, that as Man is made for Society, so he ought to be born with a kind Affection to the whole, of which he is a part, and a Propensity to seek the Welfare of it" (p. 371-372). But while these views "are a high Compliment to Human-kind", Mandeville (1714/1988, Vol. 1) says: "What Pity it is that they are not true" (p. 372).

Opposing Hobbes, Shaftesbury argued that "rational and nonrational creatures have, by nature, strong disinterested tendencies to associate with other members of their species and to promote their species' good" [66] (Maurer, 2019, p. 40). In turn, Mandeville claims that "neither the Friendly Qualities and kind Affections that are natural to Man, nor the real Virtues he is capable of acquiring by Reason and Self-Denial, are the Foundation of Society" (Mandeville, 1714/1988, Vol. 1, p. 428). One might be inclined to see the dispute between Shaftesbury and Mandeville as a dispute between psychological altruism and psychological egoism. However, the dispute between these two authors cannot be properly understood if we reduce their views to psychological altruism and psychological egoism. Like Hobbes, Mandeville argues that the foundations of society cannot be our "friendly qualities", because these are *weak*. So, Mandeville's criticism does not entail that friendly qualities *do not exist*.

I will now consider the evidence that Mandeville, in fact, accepted the existence of what we would now call altruistic ultimate desires. Although Mandeville thinks that "friendly qualities" cannot be the basis of society, he seems to accept that they exist. I say that he "seems

[66] I discuss Shaftesbury's view in more detail in the next chapter.

to accept" because it is difficult to be sure about what Mandeville claims. As Maurer (2019) says, Mandeville was not only a philosopher, but also a literary writer, and his philosophy is not the most systematic (pp. 58-59). This makes it difficult to be sure about what he believed, for his writing is often ambiguous and vague.

The first piece of evidence in favor of the view that Mandeville did not endorse psychological egoism is his discussion of *pity*. Differently from Hobbes (1640/1928), who, in his early work, offered an egoistic definition of pity (p. 30), Mandeville (1714/1988, Vol. 1) defines pity as "a Fellow-feeling and Condolence for the Misfortunes and Calamities of others" (p. 287). After offering this definition, Mandeville accepts the existence of such passion, saying that "all Mankind are more or less affected with it; but the weakest Minds generally the most" (Mandeville, 1714/1988, Vol. 1, p. 287). Mandeville does not argue that pity is derived from egoistic desires. He is more interested in making it clear that it is *not virtuous*: since pity "is an Impulse of Nature, that consults neither the publick Interest nor our own Reason, it may produce Evil as well as Good" (Mandeville, 1714/1988, Vol. 1 p. 42).

Another important view proposed by Mandeville in the second part of *The Fable of the Bees* (1714/1988, Vol. 2) is that egoistic "self-love" includes the care for one's offspring. Commenting on the order in which passions guide human action, Mandeville (1714/1988, Vol. 2) says that "[s]elf-love would first make it scrape together every thing it wanted for Sustenance, provide against the Injuries of the Air, and do every thing to make itself *and young Ones* [emphasis added] secure" (p. 138). This indicates one way in which Mandeville's egoism differs from the contemporary account of egoistic motivation. Mandeville, differently from the contemporary authors, does not seem to consider that taking care of one's offspring counts as something beyond egoism. Caring for one's neighbor's offspring could count, but caring for

one's own offspring does not. In Chapter 8, I return to this idea when I discuss the scope of altruism.

In another passage, Mandeville (1714/1988, Vol. 1) accepts the possibility that some individuals may "from no other Motive but their Love to Goodness, perform a worthy Action in Silence" (p. 41). He allows that "a few individuals in any generation eventually come to love virtue for its own sake" (Welchman, 2007, p. 59). Although affirming that they are extremely rare, it seems that Mandeville accepts the existence of these actions that we would consider non-egoistic. In another passage, Mandeville (1714/1988, Vol. 1) claims that even the virtuous must confess "a certain Pleasure he procures to himself by Contemplating on his own Worth" (p. 43). Although this seems to suggest that the motivation is, ultimately egoistic, this is not necessarily the case. Remember, as I explained in Chapter 2, one might derive pleasure from the satisfaction of non-egoistic desires without being egoist for it[67] (see Batson, 2011, p. 22).

In this section, I argued that Mandeville's moral philosophy neither relies on nor can be summarized by psychological egoism. Mandeville pictures humans as guided by self-love. As Maurer (2019) comments, "[t]oo hasty a restriction of the semantic field of 'self-love' to our contemporary notion of [psychological] egoism risks distorting our understanding of the period's debates" (p. 2). If we read the dispute between Mandeville and Shaftesbury as a dispute over psychological altruism, we are misinterpreting the debate by taking a secondary feature to be the main issue under dispute. The interpretation I defended here is accepted by some authors.

[67] However, we might question whether Mandeville's position is, in really, a moderate suspension of judgment. In the second volume of *The Fable of the Bees* (1714/1988, Vol. 2), in a dialogue between two fictitious characters, we see one of them, Horatio, asking whether one could become truly virtuous through education. The other character, Cleomenes, who represents Mandeville's views, says: "Yes, if it really was obtain'd: But how shall we be sure of this, and what Reason have we to believe that it ever was?" (Mandeville, 1714/1988, Vol. 2, p. 106).

Welchman (2007), for example, claims that "[w]hile Mandeville often asserts we act only for self-love, he does not deny that among the things and states of affairs we desire are the happiness and welfare of other individuals" (p. 63). In the next section, I address the role of egoism in Bentham's work.

6.4 Bentham and The Governance of Pleasure

As discussed in the previous chapter, the version of psychological egoism that is considered the most difficult to refute is *psychological hedonism*, which claims that "the only ultimate desires that people have are attaining pleasure and avoiding pain" (Sober, 2013, p. 150). Modern authors, such as Locke and Hobbes, consider pleasure and pain to be central to human motivation. However, it is the founder of modern utilitarianism, Jeremy Bentham (1748–1832), who is most commonly associated with this particular form of psychological egoism. The view that Bentham holds psychological hedonism remains popular (Sober & Wilson, 1998; Crimmins, 2021; Moore, 2019; Hampton, 1995; Feinberg, 2013).

In this section, however, I dispute the claim that Bentham endorses psychological hedonism. I do so by showing how we can understand Bentham's account of egoism differently, in a more parsimonious way. Bentham's hedonistic view can be explained without the need for psychological hedonism. Furthermore, following the pattern of the previous sections, I will show that there is evidence that Bentham accepted the existence of non-egoistic desires. Bentham's claim that we are always governed by hedonistic motivation is better understood as a useful generalization.

The opening lines of Bentham's most influential book, *An Introduction to the Principles of Morals and Legislation* (1789/2000), can help us see why he is associated with psychological hedonism.

> Nature has placed mankind under the governance of two sovereign masters, pain and pleasure. It is for them alone to point out what we ought to do, as well as to determine what we shall do. On the one hand the standard of right and wrong, on the other the chain of causes and effects, are fastened to their throne. They govern us in all we do, in all we say, in all we think: every effort we can make to throw off our subjection, will serve but to demonstrate and confirm it. In words a man may pretend to abjure their empire: but in reality he will remain subject to it all the while. (Bentham, 1789/2000, p. 14)

The passage above suggests a hedonistic view, where pleasure (and pain avoidance[68]) is what one values. But Bentham puts forwards two distinct kinds of hedonism here, one *descriptive* and one *normative* (Moore, 2019). The *descriptive* version of hedonism claims that hedonistic states (pleasure and pain avoidance) are what motivates our actions. The *normative* version claims that hedonistic states are what one should value ultimately. These two forms of hedonism are logically independent of each other — the truth or falsity of one does not entail the truth or falsity of the other. I will explain why these two forms of hedonism do not assume or imply psychological hedonism. Consider first the normative version.

Bentham's normative hedonism is based on the "principle of utility". This principle approves or disapproves every action "according to the tendency it appears to have to augment or diminish the happiness of the party whose interest is in question" (Bentham, 1789/2000, p. 14). The ultimate guide for moral action is to increase the net pleasure and reduce pain, not of the

[68] Pleasure and pain are to be understood in a broad sense (see Bentham, 1789/2000, p. 33, 35). In Bentham, pleasure includes many forms of positive emotions, such as joy, relief, tranquility, and so on (Moore, 2019).

agent, but that of everyone — "morality is about making the world as happy as possible" (Rachels & Rachels, 2015, p. 99). The principle of utility is what Bentham claims that legislators have to follow if they aim to promote the good of society. Bentham (1789/2000) claims that "it is but tautology to say, that the more consistently it [the principle of utility] is pursued, the better it must ever be for human-kind" (p. 22).

The idea underlying Bentham's normative account is that pleasures and pains are what we should value ultimately. He thought that different normative principles fail because they are, ultimately, imperfect attempts to promote what the principle of utility aims for directly. Consider, for example, the principle of sympathy, which Bentham (1789/2000) defines as approving or disapproving of actions based on one's subjective dispositions towards others (p. 23). The legislator, guided by the principle of sympathy, is subjected to the contingencies of his or her own dispositions, which may or may not lead to a positive outcome for others. The principle of utility is not subjected to such flaws. The principle of utility is at the base of Bentham's utilitarianism, which is one of the dominant normative theories today[69].

But from the thesis that pleasures and pains are what we should value, it does not follow that this is what we do value. So, the normative hedonism of Bentham is compatible with the existence of altruistic ultimate desires. It merely states what should be guiding our moral decision-making. It is descriptive hedonism that is considered to demand the truth of psychological hedonism.

For whatever action, says Bentham (1789/2000), "there is nothing by which a man can ultimately be made to do it, but either pain or pleasure" (p. 27). This suggests a *reduction* of

[69] Darwall (1995) explains that Bentham's normative hedonism is a polished version of utilitarian ideas developed in early modern philosophy, at least since Richard Cumberland (1631–1718) (p. 82).

one's self-interest to one's hedonistic states. A "thing is said to promote the interest, or to be for the interest, of an individual, when it tends to add to the sum total of his pleasures" (1789/2000, p. 15). When deciding what action to take, we should measure and compare the consequences of the possible actions. The individual measures the estimated value of pleasures and pains and decides what to do (Bentham, 1789/2000, p. 31). In this view, "humans are calculating machines" (Garson, 2016, p. 4), always trying to maximize their pleasure and avoid pain. This model of human motivation became very influential in economics (Clavien & Chapuisat, 2013).

A simple interpretation is that Bentham is claiming that pleasure and pain are the content of our ultimate desires: the only thing we can desire ultimately is to increase pleasure and reduce pain. If this is the case, then Bentham is assuming psychological hedonism. However, there are alternative interpretations that make hedonistic states the base of our motivation without having to make them the content of our ultimate desires.

In Chapter 3, I discussed different theories of desire. One of these theories was the hedonistic theory of desire, which claims that "[t]o desire that P is to be so disposed that one will tend to feel pleasure if it seems that P, and/or displeasure if it seems that not-P" (Schroeder, 2004, p. 27). The alternative interpretation of Bentham's hedonism that I want to propose is to take it as endorsing a hedonistic theory of desire instead of endorsing psychological hedonism. The idea is that, instead of considering Bentham's hedonism to be about the *content* of our ultimate desires, we can consider it as a claim about the *nature* of our desires. If we do so, we still keep Bentham's primacy of hedonistic states in our motivation, but we allow the existence of ultimate desires in which the content is not to pursue pleasure or avoid pain.

Accepting the view of Bentham as assuming a particular hedonistic theory of desire rather than holding psychological hedonism would mean that he is not opposing psychological

altruism. But although the alternative interpretation I propose is possible, do we have good reasons to adopt it? In the next paragraphs, I present some evidence that encourages a positive answer to this question. The alternative I propose here is more consistent with his views.

Just like Hobbes and Mandeville, Bentham also presents a set of definitions of motives to act. These motives receive a different name depending on whether they are good or bad (based on the principle of utility) (Bentham, 1789/2000, p. 85). Some "good" motives are based on what we would call egoistic desires. For example, "love of reputation" is "the desire of ingratiating one's self with, or, as in this case we should rather say, of recommending one's self to, the world at large" (Bentham, 1789/2000, p. 88). This can lead to actions that benefit others. So, moral actions do not depend on altruistic desires per se. Bentham's utilitarianism can account for good deeds without demanding humans to desire the good of others ultimately. Nevertheless, Bentham is clear when he says that these other-regarding motives *do exist*.

In his list of motives, Bentham mentions the motive of "good-will", which he defines as "the pleasure of sympathy" (1789/2000, p. 90). When good-will is good, it is called "*benevolence*"[70]. He explains that "[t]he pleasures of benevolence are the pleasures resulting from the view of any pleasures supposed to be possessed by the beings who may be the objects of benevolence" (Bentham, 1789/2000, p. 37). The motive of benevolence seems to produce pleasure when we see people worthy of pleasure to obtain it. This does not require an inferential process where we infer benevolent desires from egoistic desires. Bentham seems to accept direct desires for the good of others. Of course, he talks in terms of pleasures that benevolence causes in us, but, if we adopt the hedonistic theory of desire, then this is a *feature* of desires, not the

[70] Benevolence is the motive Bentham believes to be the closest to the principle of utility itself. What the principle of utility promotes is, basically, what benevolence would promote in an epistemically "enlightened" context (Bentham, 1789/2000, p. 97).

content of our ultimate desires. In another passage, Bentham is more direct in pointing out that some desires are directed at others.

> [T]here are certain pleasures and pains which suppose the existence of some pleasure or pain, of some other person, to which the pleasure or pain of the person in question has regard: such pleasures and pains may be termed extra-regarding. (Bentham, 1789/2000, p. 40)

The passages above indicate Bentham's account of human motivation as not strictly based on egoistic ultimate desires. In another text, where Bentham discusses the idea of assuming people as pursuing their own self-interest, he says that even though the idea that all of our motivations are self-directed "*held good in no more than a bare majority* [emphasis added], of the whole number of instances, it would suffice for every practical purpose, in the character of a ground for all political arrangements" (Bentham, 1843, p. 6). That is, the idea that we are motivated only by self-regarding desires is only a *generalization* — there are exceptions to it, but, since it is so pervasive, we are justified in holding it as the standard view (see Crimmins, 2021).

Bentham, who was also a philanthropist, recognized and valued altruistic actions. His philosophy, although giving pleasure a central role in our motivation, allows that these altruistic actions can be motivated by genuine "extra-regarding" motives. We can account for Bentham's hedonism without relying on the metaphysically extravagant hypothesis of psychological egoistic hedonism. The next section discusses further ways in which we can understand the "egoistic" view of Bentham without relying on psychological hedonism.

6.5 Many Ways of Loving Oneself

In his book *Self-love, Egoism and the Selfish Hypothesis* (2019), Christian Maurer discusses how egoism, selfishness, and the 18th-century term "self-love" can refer to a plurality of phenomena. Maurer alerts that present-day readers might be tempted to promote a simplistic interpretation, reducing these motivations to the contemporary accounts of egoistic motivation. To navigate the plurality of egoisms, Maurer (2019) proposes *five* different accounts of self-love present in 18th-century British philosophy: "1) self-love as egoistic or self-interested desires, 2) self-love as love of praise, 3) self-love as self-esteem (or due pride), 4) self-love as amour-propre (or excessive pride) and 5) self-love as respect of self" (p. 3). As he comments, it is "essential to distinguish the five aforementioned conceptions of self-love, and to resist too hasty an assimilation of self-love with our present-day concept of egoism" (Maurer, 2019, p. 12).

Along with Maurer (2019), I am interested in showing how we should resist a hasty interpretation of modern authors in terms of contemporary concepts of egoism and altruism. In previous sections, I discussed how some modern accounts of egoism differ from psychological egoism. In this section, I go further and propose some alternative explanations for these egoistic accounts that are not dependent on and do not entail psychological egoism. In addition to supporting the claim that we should disassociate modern authors from psychological egoism, this section also aims to offer alternatives to readers who are inclined to view egoism as the ultimate source of motivation but are reluctant to accept psychological egoism. Believing in the primacy of egoism does not entail that one should endorse psychological egoism, and here I discuss alternative egoistic accounts.

As I explained, the emphasis that Hobbes gives to egoistic motivations can be understood as a matter of the *relevance* of these motivations for his argument. Considering this, Kavka (1986) argues that, rather than psychological egoism, Hobbes holds *predominant egoism*, which

states that "self-interested motives tend to take precedence over non-self-interested motives in determining human actions"[71] (p. 64). The reason why Hobbes does not give much attention to non-egoistic desires is not that he thinks that they do not exist, but simply that they are weak and not relevant to his political argument. Predominant egoism does not entail the rejection of altruistic ultimate desires, so it is compatible with psychological altruism. It merely states the superior strength of egoistic desires.

If we interpret Hobbes as assuming predominant egoism, we can easily explain how he could accept the existence of altruistic motivation. Authors such as Hampton (1995) see the reading of Hobbes as assuming predominant egoism as more parsimonious than reading him in terms of psychological egoism. As Hampton (1995, p. 22) claims, "we should not attribute to him a crude and probably false psychological view of all human motivation [psychological egoism] when his text suggests a better alternative".

Maurer's (2019) five self-loves are particularly helpful to assess Mandeville's views on egoism. In Mandeville, self-love is what we call egoistic desire, such as the desire for self-preservation. Mandeville's account of self-love, as a self-interested desire, fits Maurer's *first* self-love. But Mandeville (1714/1988, Vol. 2) proposes a different account of egoism, which he calls "self-liking" (p. 134). Contrasting with self-love, self-liking is "an Instinct, by which every Individual values itself above its real Worth" (Mandeville, 1714/1988, Vol. 2, p. 134). Self-liking fits the *fourth* self-love in Maurer's list, which is *amour-propre* or excessive pride (Maurer, 2019, p. 61). Considering that Mandeville considers self-liking the fundamental motivation underlying our social virtues, Maurer (2019) argues that we should prefer the

[71] Predominant egoism is similar to what Sober and Wilson (1998) call "E-over-A Pluralism" (p. 245). In this view, egoistic desires are dominant, while altruistic desires, although existent, are weaker.

interpretation of Mandeville's egoism as *amour-propre* or excessive pride rather than psychological egoism: in Mandeville, "*amour-propre* plays the central role, rather than self-love as egoistic desire" (p. 70).

Finally, Bentham's hedonism can also be accounted for without relying on psychological hedonistic egoism. I already offered the interpretation of Bentham's hedonism as a theory of desire. But here I will consider yet another account of egoism that can be used to interpret Bentham's views, which is proposed by Garson (2016).

Developing ideas present in LaFollette (1988), Garson (2016) proposes an alternative descriptive hedonism, which he calls "*reinforcement hedonism*". Reinforcement hedonism "holds that, where D is an ultimate desire, D is maintained or reinforced in A's cognitive system only by virtue of the fact that D is associated with pleasure" (Garson, 2016, p. 1). Different from psychological hedonism, reinforcement hedonism is not a theory about the *content* of desires but about the "mechanism by which that desire is reinforced in the cognitive life of the agent" (Garson, 2016, p. 2).

Reinforcement hedonism is an alternative to psychological hedonism. Garson (2016) claims that, differently from psychological hedonism, reinforcement hedonism is supported by neuroscientific evidence (p. 10). Relying on the empirical work of Dickinson and Balleine (2010), Garson argues that reinforcement hedonism is supported by the neuroscientific understanding of pleasure.

> [P]leasure need not regulate human behavior by being part of the content of human desires. It need not be the thing we are explicitly aiming for. Rather, the function of pleasure is to provide a kind of reinforcement mechanism that strengthens or weakens the values associated with certain goals. Dickinson's work shows that pleasure need not be part of the content of a desire in order to serve as a reinforcement mechanism for our desires. (Garson, 2016, p. 11)

Reinforcement hedonism "borrows its strength from the sorts of intuitions that have always sustained [psychological hedonism], and avoids all of its weaknesses" (Garson, 2016, p. 15). It accounts for the role of pleasure in desiring, giving an alternative to psychological hedonism. Both the hedonistic theory of desire and reinforcement hedonism are examples of how we can account for Bentham's hedonistic motivational account without having to make the strong claim that the content of all of our ultimate desires is pleasure and pain. Neither a hedonistic theory of desire nor reinforcement hedonism is inconsistent with psychological altruism (Garson, 2016, p. 9).

A final challenge to the interpretation of these three authors as proponents of psychological egoism is the distance between their accounts of motivation and our contemporary way of characterizing egoistic motivation. It is not clear to what degree it is reasonable to hold that these 17th- and 18th-century authors were discussing what we now call "ultimate desires". They talk about "passions", "affects", "sentiments", and so on. Sometimes interchangeably, sometimes to describe different things (Schmitter, 2021). It is difficult for us to affirm that they are talking about what we call ultimate desires and not another motivational state, such as emotions.

As Maurer (2019) shows, modern philosophy offers a rich repertoire of ideas associated with egoism, which might be rashly interpreted in terms of the contemporary notion of egoistic motivation. In this chapter, I have shown that reading modern authors in terms of psychological egoism is problematic. Remember that the strong historical assumption, mentioned in the first section, says that psychological egoism and psychological altruism are the central issues under dispute in the modern discussion on egoism. The conclusion I reach is that this strong version of the historical assumption should be rejected. Psychological egoism does not play a central role in

the philosophy of the authors who are often assumed to defend it. We have evidence that these authors even accepted the existence of altruistic ultimate desires. Reading them through the lens of psychological egoism distorts and impoverishes their views. Furthermore, the alternative forms of egoism discussed here offer viable alternatives for individuals who are intuitively inclined to a more self-centered account of motivation. One should not be asked to choose between psychological egoism or accepting psychological altruism: one can reject both of them as bad ways of thinking about human motivation.

The strong historical assumption portrays a rich historical foundation for the contemporary debate on psychological altruism, playing an important role in the popularity and legitimacy of this debate. If it is true that psychological altruism and psychological egoism have been the focus of prominent philosophers for centuries, readers are likely to give more credit to the contemporary debate. The "altruism question", then, is not an ordinary research project that should be abandoned if it proves to be unfruitful, but an inquiry into a fundamental aspect of human nature. By undermining this historical assumption, this chapter adds a further reason for us to challenge the contemporary approach to altruistic and egoistic motivations.

In the next chapter, I move on to the history of *altruism*. There, I will give further reasons why we should reject the historical assumption in its strong form. One of these reasons, which I did not explore in this chapter, is how the contemporary debate on psychological altruism tries to be independent of the *moral* dimension of altruism and egoism. When we interpret the modern debate as a debate over psychological altruism, we dismiss the moral dimension of this debate. However, as Maurer (2019) comments, for modern authors "the psychological dimension is conceived as inextricably connected to the moral one" (p. 3). The next chapter shows how the

notions associated with altruism in modern philosophy, as well as the original account of altruism in Comte's philosophy, all have a normative dimension that cannot be ignored.

Chapter 7

The Origins of Altruism

7.1 The Long History of Caring for Others

Many concepts in the history of philosophy can be interpreted as identifying or pointing towards forms of altruism insofar as they describe certain motivations that are in some sort of opposition to egoistic motivation. But to what degree can we interpret these historical concepts in terms of ultimate desires to increase the welfare of others? Is our contemporary account of altruistic motivation a good representative of the motivation that authors throughout history have proposed as the opposite of egoism and selfishness? This chapter discusses some historical accounts of "anti-egoistic" motivation, comparing them to the contemporary account of altruistic motivation.

As a starting point, consider a few ancient philosophers and how they thought about motivations to help others. One of the earliest philosophical discussions about dispositions to care for others dates back to the 6th century BCE, in the Chinese philosophical tradition of Confucianism. The concept of "*ren*" represents a central virtue in the ethical teachings of Confucius (551–479 BCE). In the *Analects*[72], which is regarded as the most accurate representation of Confucius's thought (Csikszentmihalyi, 2020), Confucius describes *ren* in different ways. In a passage, he defines *ren* simply as "loving others" (Brooks & Brooks, 1998,

[72] Translated and edited by Brooks and Brooks (1998).

p. 95). Scholars have translated *ren* in different ways, including "benevolence" and "humanity", which indicates proximity to the idea of altruism[73] (see Dubs, 1951, p. 48).

In the paper "The Development of Altruism in Confucianism" (1951), the Sinologist Homer Dubs discusses the association between *ren* and altruism. Although loving others is central to the Confucian virtue of *ren*, this is not a universalized love for others. Confucius recognized that people tend to care more for their relatives than for strangers. Rather than criticizing such a bias, Confucius accepted it as correct and natural (Dubs, 1951, p. 49). In Confucius, the love for others was (and should be) graded, being stronger for close relatives and weaker for unknown people. Furthermore, *ren* was not only supposed to be directed to a few, but also to be *performed* by a few: servants and individuals from lower classes should cultivate other virtues, such as loyal obedience, while *ren* was more like "the attitude of a bountiful lord to his inferior" (Dubs, 1951, p. 49).

Confucius's exclusionary conception of "love for others" contrasts with that of another important Chinese philosopher, Mozi (470–391 BCE), who founded the movement of Mohism. Like Confucius, Mozi believed that loving others is a central moral tenet. However, Mozi criticized Confucius's view that this love should be directed primarily toward one's close kin. As Dubs (1951) comments, Mozi mentions that many vicious activities can follow from loving one's own family without also caring for others: "[t]he thief and the aggressor are not without any love. Their love is merely restricted. Because their loves are graded so sharply... they rob or attack other persons" (p. 50). Mozi argued that we should make our love for others more inclusive, and he expressed this idea in the concept of "*jian ai*", which can be translated as

[73] Ren is also considered to represent, more broadly, a chief virtue. Confucius says that "to overcome the self and turn to propriety is ren" (Brooks & Brooks, 1998, p. 89).

"universal love" or "inclusive care" (Fraser, 2022). The scope of our care for others should not be restricted by the limits of our kinship bonds, but extended to include the whole of humanity[74].

Moving to the Western tradition, one can find Plato (428/427–348/347 BCE) discussing some ideas that are foundational to our way of thinking about altruism and egoism. For example, in Plato's *Meno* (1997, pp. 870-897) we find the idea of interpreting desires through the "hunger model". The starting point of this model is to assume that, when we are hungry, we want food only for the sake of the satisfaction it causes. The model generalizes this pattern to all desires: all things we desire we desire for some sort of satisfaction they cause in us. Thus, assuming this model, we do not desire anything for its own sake, but only as a means to the satisfaction of an egoistic desire[75] (see Kraut 2020). This way of thinking about desires in an a priori way influenced much of the later defenses of psychological egoism (see Feinberg, 2013).

Aristotle (384–322 BCE) discusses the idea of caring for others for their own sake more directly. This is central in his discussion of friendship, in the Book VIII of his *Nicomachean Ethics* (Aristotle, 2000, pp. 143-163). There, Aristotle distinguishes three kinds of friendships. The first two are based on the individuals' expectations of getting something out of the friendship, either some advantage or pleasure. But in the third kind of friendship, which can only occur between virtuous individuals, one wishes the good of the other *for their own sake*

[74] Although Mozi's *jian ai* was vehemently rejected by traditional Confucianism as something that goes against nature, it was eventually accepted and integrated into the Neo-Confucianism tradition, centuries later (Dubs, 1951, p. 53).

[75] However, as MacIntyre (1967) explains, for Plato, "the pursuit of *good as such* and the pursuit of *my good* necessarily coincide" (p. 463). So, although he introduced ideas such as the hunger model, Plato was not defending an egoistic view of human nature.

(Aristotle, 2000, pp. 147, 149). We can see in Aristotle the germ of the idea of a genuine concern for others as opposed to an interested form of concern for others, which today we call egoistic[76].

As the brief discussion above illustrates, the idea of caring for others without expecting rewards is present in philosophy since its origins. Today, we have a particular way of thinking about this issue. The debate is framed in terms of the opposition between the view that all of our motivations are ultimately egoistic and the view that at least some of our motivations are ultimately altruistic. But the contemporary approach is only one way of thinking about the idea of motivations to care for others. There are alternative approaches. By looking at the ancient philosophers mentioned above, we can already notice that our contemporary concepts might not do a very good job articulating their views. This chapter discusses other influential authors that proposed some sort of non-egoistic motivation and investigates to what degree the standard account of altruistic motivation can represent their views.

In the previous chapter, I discussed how early modern philosophy is a particularly important period for the development of ideas regarding egoistic motivation. Ancient and Medieval philosophers, particularly in the Aristotelian-Thomistic tradition, could rely on the idea of a "perfectionist/teleological metaphysics" (Darwall, 1995, p. 5). This metaphysical foundation guarantees that people following their own (true) self-interest will end up producing a world where everyone is better off. Egoistic motivation, thus, would not be a problem. But without this metaphysical guarantee, authors such as Hobbes had to deal with the conflict between egoistic motivation and social harmony. Egoism was a much more pressing problem for modern

[76] The idea of considering the roles of beliefs and desires as we do today can be tracked back to Aristotle. Kahn (1987) argues that we do not find a similar approach to beliefs and desires in previous philosophers, such as Plato (p. 78). The distinction between what we want instrumentally and ultimately is also present in Aristotle (2000) in a way that is similar to how we think of these today (p. 9).

philosophers than it was for ancient and medieval philosophers. Considering this and the fact that I discussed only modern philosophers in the previous chapter, the next section is dedicated to modern philosophers. I will address some authors from the tradition of British Moralism that, differently from the authors discussed in the previous chapter, argued that the basis of civility and morality are our non-egoistic dispositions.

Although we can find in modern philosophy the idea of motivations that we can call "altruistic", the term "altruism" was not part of their repertoire. The term "altruism" did not exist before the 19th century. The French word "*altruisme*" first appeared in 1851, in the writings of the founder of Positivism, Auguste Comte (1798–1857). One year later, the English translation "altruism" was introduced by the philosopher G. H. Lewes (Dixon, 2008, p. 1). Since altruism was born in Comte's philosophy, it is imperative for any dispute about the meaning of altruism to have a clear understanding of what he meant by altruism. Thus, much of this chapter will be dedicated to Comte's work. In *Section 7.3*, I present an introduction to Comte's positivism, explaining the context in which the idea of altruism was originally conceived. *Section 7.4* narrows the focus to Comte's understanding and categorization of mental functions. I explain his "cerebral theory," based on which the idea of altruism is articulated.

Finally, *Section 7.5* discusses the way in which altruism was assimilated into the scientific discourse of the 19th century, particularly in evolutionary theory. It starts by covering Comte's evolutionary views and then discusses the work of Herbert Spencer (1820–1903), who also used altruism as a key concept in articulating a naturalistic account of ethics. If Comte introduced altruism to the world, it is in Spencer's work that we find the roots of evolutionary altruism. His behavioral account of altruism, and the proximity of such an account to Comte's views, are important chapters in the history of altruism. In passing, I shall also discuss Darwin's

evolutionary morality, showing how some of his views on the evolution of our non-selfish instincts were similar to Comte's.

7.2 The Pre-History of Altruism: Social Affections in Modern Philosophy

This section focuses mainly on the work of Anthony Ashley Cooper, the Third Earl of Shaftesbury (1671–1713), who started a tradition that emphasized the existence and importance of non-egoistic psychological traits in human nature. He is regarded as "the thinker who set the stage for the debates on self-love in eighteenth-century British moral philosophy" (Maurer, 2019, p. 32). I will also discuss some ideas of Francis Hutcheson (1694–1746), who refined some of Shaftesbury's ideas into a more sophisticated version and heavily influenced the modern debate. Finally, I also address Bishop Joseph Butler (1692–1752), who is often considered to be the author that has shown that psychological egoism is wrong.

One of the main problems for moral philosophers of the early modern period was finding the basis for the normative power of morality (Crisp, 2019). How can we be motivated (and obligated) to behave morally? Answering this question in this period was particularly challenging for there was "a widespread belief that any defensible moral philosophy, including any acceptable account of moral obligation, must be consistent with modern science, if not itself scientific" (Darwall, 1995, p. 7). One of the most influential early modern naturalistic moral accounts was that proposed by Hobbes. As discussed in the previous chapter, Hobbes explains how individuals pursuing their own self-interest can be led to abandon the state of nature and live in civil society. But as the Hobbesian approach received more and more attention, it also

attracted critics. Some of Hobbes's contemporaries protested against what they believed to be an overly selfish account of human nature. One of these critics was Shaftesbury.

Shaftesbury's philosophy aims to free morality from self-interest. To understand his argument, we need first to understand the notion of "moral sense", which was very influential in modern philosophy (Gill, 2021). This idea was introduced by Shaftesbury in his *An Inquiry Concerning Virtue, or Merit* (1711/2001, Vol 2., pp. 3-100) and received a more polished version later in Hutcheson's *An Inquiry Concerning the Original of our Ideas of Virtue or Moral Good* (1725/2004, pp. 83-197). The moral sense indicates what is morally praiseworthy and what is morally blameworthy. It distinguishes virtue and vice analogously to how an aesthetic sense distinguishes what is harmonious and disharmonious. As Hutcheson (1725/2004) explains, a sense is the "determination of the mind, to receive any idea from the presence of an object which occurs to us, independent of our will" (p. 90). The moral sense works independently from our personal preferences. So, if our moral approbations stem from our moral sense, that means that they cannot be based on our personal preferences any more than the images we see and the sounds we hear can be determined by our preferences. The moral sense indicates the *independence* between morality and our self-oriented preferences.

Shaftesbury (1711/2001, Vol. 1) claims that, on top of having a moral sense, humans are also naturally endowed with disinterested tendencies, which he calls "social" or "natural" affections. Rejecting the view that "self-interest governs the world", he states that "Passion, Humour, Caprice, Zeal, Faction, and a thousand other Springs, which are counter to Self-Interest, have as considerable a part in the Movements of this Machine" (Shaftesbury, 1711/2001, Vol. 1, p. 115). For Shaftesbury, social affections are present in our motivational repertoire. More importantly, they are *as strong as* (or even stronger than) the opposing selfish affections

(Shaftesbury, 1711/2001, Vol. 1, p. 116). The strength of the social affections is the crucial point distinguishing Shaftesbury from authors such as Hobbes and Mandeville.

To see the importance of social affection, we have to consider that the moral sense is not directed to *actions* but to *motivations* (Kauppinen, 2022). The motivations that our moral sense approves and views as praiseworthy are precisely the social affections. Shaftesbury believed that the concepts of moral sense and social affection offer the basis for an account of morality free from self-interest.

For Shaftesbury (1711/2001, Vol 2.), in order for an affection to be approved by the moral sense, it needs to have the right scope: such an affection needs to be suitable with "the Good of his Kind, or of that System in which he is included, and of which he constitutes a PART" (p. 45). More than mere desires to benefit others, social affections are motivational states that aim at *the good of the species* (Maurer, 2019, p. 40) or, more broadly, at the good of the system composed of all rational creatures (Gill, 2021).

Hutcheson (1725/2004) echoes Shaftesbury's views explaining that "the Actions we approve in others, are generally imagin'd to tend to the natural Good of Mankind, or of some Parts of it" (p. 91). Concerned with the scope that our concern for others should have, Hutcheson (1725/2004) claims that "all those kind Affections which incline us to make others happy… appear morally Good, if while they are benevolent toward some Persons, they be not pernicious to others" (p. 116). An important theoretical resource helping Hutcheson here is his classification of different *kinds of benevolence*. He considers two variables, namely whether benevolence is calm or violent, and the scope of such benevolence (Maurer, 2019, p. 89). It is calm universal benevolence that is regarded as the superior form of benevolence (Hutcheson, 1728/2002, p. 32). If benevolence is violent or has an excessively narrow scope, then it is unlikely to be approved

by our moral sense. If we read Shaftesbury and Hutcheson as mere proponents of psychological altruism, this complex dimension regarding the scope and the morality of our concern for others would be neglected.

As we can see, the motivational state of focus in the contemporary account of altruistic motivation would not be necessarily approved by the moral sense. A desire to improve the welfare of just any other would often be directed toward the wrong targets. Another reason for not considering altruistic motivation as something that would be approved by the moral sense is that, for Shaftesbury, the simple *existence* of a social affection is not a sufficient condition for one's action to be approved by the moral sense. To be approved, a social affection needs to actually cause a behavior. Shaftesbury (1711/2001, Vol 2.) explains that, in some individuals, the natural affections are present but with "insufficient force", and if such affection "is not in its natural degree, 'tis the same indeed as if it were not" (p. 59). That is, the effective influence over one's behavior is a condition for a social affection to be approved. Remember that, in psychological altruism, such a condition is not required: even if an ultimate altruistic desire does not produce any behavior, its mere existence already makes psychological altruism true.

A question we can ask is whether social affections can be equated with altruistic motivation in the contemporary sense of the term. I will argue that there are at least *two* reasons for responding to such a question in the negative. The *first* reason is that the "self-interest" that Shaftesbury is opposing is not the abstract entity we call ultimate egoistic desire, but something that I will call "deliberate egoism". The *second* reason is that it is not clear that social affections are (or assume) ultimate altruistic desires: they might be ultimately grounded on what we would call egoistic desires.

Shaftesbury's moral philosophy rejects a conception of morality based on egoism or self-interest. But the egoism he was targeting was not psychological egoism, but the idea that the foundation of the human heart is a "cool and deliberate selfishness" (Shaftesbury, 1711/2001, Vol. 1, p. 116). As Maurer (2019) comments, Shaftesbury is objecting to "the claim that we always act upon a rational assessment of how best to promote our interest" (p. 33). The self-love which he sees as pernicious is not merely a motivation aiming to benefit ourselves, but the "steddy [sic.] and deliberate Pursuit of the most narrowly confin'd Self-interest" (Shaftesbury, 2001, Vol 2., p. 46). Such "deliberate egoism" obviously differs from psychological egoism. Ultimate egoistic desires do not need to be conscious nor to be accompanied by a calculation of benefits and costs[77]. Therefore, the egoism rejected by Shaftesbury should not be equated with psychological egoism.

Consider now the second reason for resisting equating social affection with altruistic motivation. Shaftesbury rejected deliberate egoism as the basis for morality. But is he considering that the social affections that replace such egoism are (or assume) ultimate altruistic desires? A first step to answering this question is to consider that Shaftesbury believed that God has assembled human beings with certain tendencies to be social and virtuous. These tendencies work by disposing us to *enjoy* helping others. Shaftesbury (1711/2001, Vol. 2) explains that cultivating our social affection is "the only means which can procure him [humans in general] a constant Series or Succession of the mental Enjoyments" and "a certain and solid Happiness" (p. 58). Here we see Shaftesbury following the Aristotelian tradition, which considers that caring for

[77] Hutcheson claims that such "deliberate selfishness", which depicts us as "self-interested calculators", depends on an unplausible "over-intellectualisation" of humans (Maurer, 2019, p. 93). Hutcheson thought that one can find counterevidence for such a form of egoism not only in the existence of social affections, but also in the irrational and violent passions, which do not follow from a rational egoistic calculation.

others for their own sake is constitutive of happiness. The question is whether these social affections are conditional to our pursuit of happiness — which is an egoistic motivation in the contemporary approach.

Although there is no consensus among scholars, the view of Shaftesbury as holding an ultimately egoistic account of moral motivation is accepted by some commentators (see Gill, 2021). In the paper "Shaftesbury's Egoistic Hedonism" (2010), Simon Grote argues that, although Shaftesbury rejects the idea that one's reason to be virtuous is a pursuit of rewards, he does not exclude "internal" rewards. For Grote (2010), Shaftesbury accepted that the ultimate reason for being virtuous is that this is the best way of pursuing one's own happiness (p. 138). Maurer (2019) has a similar interpretation, claiming that Shaftesbury "points towards an egoistic reason for realising our nature by cultivating natural sociability" (p. 36). If Grote and Maurer are correct, then the standard account of altruistic motivation would not only neglect some important aspects of Shaftesbury's philosophy but would straightforwardly misrepresent his views.

Shaftesbury did not seem to be very concerned with being regarded as a proponent of some ultimate form of egoism. He says that, if we address properly the issue of whether we are motivated by self-interest, we would not ask "Who lov'd himself, or Who not", but instead "'Who lov'd and serv'd himself the rightest, and after the truest manner'." (Shaftesbury, 1711/2001, Vol. 1, p. 121). The goal, for him, is achieving the "right balance of several kinds of affections" (Maurer, 2019, p. 37). Shaftesbury (1711/2001, Vol. 1) goes as far as to say that "the height of Wisdom" is to be "Rightly selfish" (p. 121).

To make sense of Shaftesbury's view, we need to remember that Shaftesbury assumes a harmonious metaphysical foundation for reality, which I explained in the previous chapter. For him, "the promotion of the private and the public good are not in conflict, but connected thanks

to the harmonious design of the universe" (Maurer, 2019, p. 43). Pursuing the correct, divinely provided self-interest would lead everyone to be better off.

The last modern author that I will discuss in this section is Bishop Joseph Butler, who raised an influential argument against the idea that every desire aims for something internal and egoistic. Butler (1726/2006) starts the argument by claiming that if there was not a desire for food, which is an external thing, then eating it would not produce pleasure. If we did not desire external things, explains Butler, there would be no reason for some things to give pleasure and others not. "There could be no enjoyment or delight for one thing more than another, from eating food more than from swallowing a stone, if there were not an affection or appetite to one thing more than another" (Butler, 1726/2006, p. 111).

Butler's argument indicates that some of our desires are directed toward external things, and this might include the welfare of other people. Some authors have considered this a case for psychological altruism (e.g., Kitcher, 2011, p. 20). However, in the paper "Hedonism and Butler's Stone" (1992), Sober convincingly argued that Butler's argument does not reject psychological egoism. By mentioning that we often desire external things, Butler rejects the hypothesis that *all* of our desires are desires for something internal, such as pleasure. This is not controversial and is accepted by both sides of the contemporary debate on psychological altruism. As Sober (1992) explains, however, Butler does not offer justifications for why these desires for external things cannot be instrumental to egoistic desires, which is what is at stake when we discuss psychological egoism.

In more recent work, Sober (2013) questions why so many people have considered Butler's argument as a case against psychological hedonism. He considers this surprising, since Butler (1726/2006) talks about how our moral behavior, ultimately, depends on the idea that it

leads to our happiness (p. 117). Sober (2013) questions whether it is fair to criticize Butler for failing to reject psychological hedonism. He asks whether the famous rejection of psychological hedonism, attributed to Butler, fails because Butler never really tried to reject such a hypothesis in the first place (Sober, 2013, p. 155).

This section shows some ways in which the views of modern philosophers diverge from the contemporary debate on altruistic motivation. In the discussion of Shaftesbury, I defended the view that social affections are not (or do not depend on) ultimate altruistic desires. The views of the other authors addressed here, Hutcheson and Butler, also diverge from an interpretation in terms of our contemporary terminology. The next section moves on to the discussion of the origins of the concept of altruism.

7.3 Comte and The Genesis of Altruism

The previous section discussed what I called a "pre-history" of altruism. I used the term pre-history because early modern philosophers did not use the term "altruism", which was invented much later, in the middle of the 19th century. In this section, I present the fascinating positive philosophy of Auguste Comte. This analysis of Comte's views allows us to better understand altruism in its genesis.

Comte's *magnum opus* was his *Course on Positive Philosophy* (*Cours de Philosophie Positive*)[78]. In this extensive work, Comte establishes the foundations of the Positivist

[78] The first volume of this work was published in 1830, and only 12 years later Comte published the last volume. The six volumes were translated by Harriet Martineau into a shorter version, composed of three volumes, called *The Positive Philosophy of Auguste Comte* (Comte, 1853/2000), which I will use here.

philosophy. A central idea in the *Course on Positive Philosophy* is the *Law of the Three Stages*, which Comte considered to be a fundamental law underlying the development of humanity (Bourdeau, 2022). This law establishes three stages of development, the *theological*, the *metaphysical*, and the *positive*, through which humanity has to pass in its progress (Comte, 1853/2000, Vol. 1, p. 27).

In the *theological stage*, the human mind seeks absolute knowledge "and supposes all phenomena to be produced by the immediate action of supernatural beings" (Comte, 1853/2000, Vol. 1, p. 28). In the *metaphysical stage*, instead of supernatural entities, explanations of phenomena are given in terms of abstract entities. This is merely a modification of the theological stage and a preparation for the third stage. Comte (1853/2000, Vol. 1) explains that this intermediate stage is necessary because the "human understanding, slow in its advance, could not step at once from the theological into the positive philosophy" (p. 30). Finally, after these two stages, humanity can reach the *positive stage*, in which the pursuit of the ultimate causes is dismissed, and scientific knowledge is the ultimate guide.

One of the important goals of Comte's positivism was to avoid the fragmentation of scientific knowledge. For him, such fragmentation restrains scientific progress and, consequently, human progress. In the *Course on Positive Philosophy*, Comte proposes a classification of the sciences. This classification is one of his best-known ideas today. He proposes a hierarchy in which sciences are arranged "according to the decrease in generality and increase in complication" (Comte, 1853/2000, Vol. 1, p. 34). The first item on the list is mathematics, followed by astronomy, physics, chemistry, and biology. This order captures not only a hierarchical order of the objects of study of each science, but also "the logical connections between the various sciences and their historical succession" (Guillin, 2018, p. 136). After these

five sciences, Comte includes the last science, a science that is dependent on others and, at the same time, responsible for coordinating them into a coherent whole. This science was the study of society, or "social physics" (*physique sociale*), which Comte later called "sociology" (*sociologie*) (Guillin, 2018, p. 128). Among the many neologisms proposed by Comte, "sociology" is arguably the most famous.

For Comte, the positive stage was already manifested in some areas of knowledge. He explains that "different kinds of our knowledge have passed through the three stages of progress at different rates" (Comte, 1853/2000, Vol. 1, p. 31). In natural sciences, for example, we can see manifestations of the positive stage through the guidance of figures such as Bacon and Galileo, who rejected the superstitious and scholastic systems (Comte, 1853/2000, Vol. 1, p. 32). The study of society, however, was still under the rule of the theological and metaphysical principles. In Comte's view, such a delay for sociology to reach the positive state follows from the fact that its object is the most complex. Comte's project aims to emancipate society, develop scientific laws of human organization, and bring our thinking about society to the positive stage. Once the positive stance adopted towards natural sciences could be adopted towards sociology, then the transition to the positive stage would be complete (Comte, 1853/2000, Vol. 1, p. 33).

The *Course on Positive Philosophy* became very popular shortly after its publication, being praised by John Stuart Mill and many other important intellectuals of the time (Mill, 1865/1969, p. 263; see Dixon, 2008, p. 42). Unfortunately for Comte, however, the same was not true for his following major work, *System of Positive Polity* (*Système de politique positive*) (1851-1854/1875-1877). In this late work, after recognizing that the progress of humanity concerns both intellectual and moral improvement, Comte moves beyond intellectual matters and engages with the moral dimension of human life. While science could emancipate our intellect,

morality was still under the governance of theological notions, mainly that of the Catholic Church. In the *System of Positive Polity,* Comte proposes a positivist foundation for morality, and altruism is a key concept in such a project.

The main goal of positivist morality is "to make our sympathetic instincts preponderate as far as possible over the selfish instincts" (Comte, 1851/1875, Vol. 1, p. 73). For Comte, the main problem of human life is the *predominance of egoism over altruism*. He complained that the Catholic church has perpetuated a very negative view of humans, in which altruistic tendencies are alien to our nature, only obtained through "superhuman Grace" (Comte, 1851/1875, Vol. 1, p. 73). Nevertheless, although the traditional religion has its flaws, Comte thought that moral practice should not be guided by intellect alone, because when left alone, intellectual development leads to an expansion of egoistic instincts. For Comte, a purely intellectual approach to morality was bound to fail. His solution to this problem is found in a central idea in the *System*, namely, the *subjective* principle of Positivism. According to this principle, the mind was supposed to be a *servant to the heart*. More precisely, "the intellect should devote itself exclusively to the problems which the heart suggests" (Comte, 1851/1875, Vol. 1, p. 15). In the next section, I return to this and explain Comte's understanding of the heart and the intellect in detail. For now, let me explain how he thought we could avoid an egoistic and purely intellectual morality.

From 1817 to 1824, Comte served as the secretary of Henri de Saint-Simon (1760–1825) (Bourdeau, 2022). At this time, intellectuals such as Saint-Simon were interested in ways of organizing post-revolutionary societies. One of the dilemmas faced by them was whether societies should have a religious element at their base or whether they should be based on self-interest alone. Rejecting both alternatives, Saint-Simon argued that we should aim at a new

religion, a "new Christianity". These views are reflected in Comte's late work. Comte rejected morality based on egoism or in traditional religion. At the same time, he considered that moral practice needed some form of religious activity: "[n]othing can take the place of a special and sustained cultivation of Universal Affection, the only internal spring of true Religion" (Comte, 1852/1875, Vol. 2, p. 42). In fact, he thought that any social reform without a "spiritual power", which gives people unity of belief, was condemned to failure (Comte, 1851/1875, Vol. 1, p. 65). So, Comte's grandiose project in the *System of Positive Polity* includes nothing less than proposing a whole new religion — a religion without metaphysics, a religion without conflicts with science, a religion without a God. To this religion, Comte gave the name "*Religion of Humanity*".

Comte's new religion was a non-intellectual and non-theological way of inspiring moral progress, promoting "the feelings of venerative, identificatory and devotional love towards Humanity" (Wernick, 2001, p. 4). The Religion of Humanity aimed to "strengthen the altruistic impulses seen as vital for the correct orientation of thinking and acting" (Wernick, 2001, p. 4). Such a new religion, for Comte, not only was "in the true sense of the word, a Religion" but "one more real and more complete than any other, and therefore destined to replace all imperfect and provisional systems resting on the primitive basis of theology" (Comte, 1851/1875, Vol. 1, p. 265). In his religion, Comte (1851/1875, Vol. 1) proposes that rather than living for God, we should live for *humanity* — a goal expressed in the positivist motto "to live for others" (p. 263).

However, perhaps unsurprisingly, Comte's ambitious project was not as successful as he envisioned. Comte went from being an influential and respected philosopher, who produced a sober philosophy aligned with the worldviews of naturalistic philosophers and scientists, to being the self-proclaimed high priest of a religion he invented. Far from substituting Catholicism,

"Comte's founding religious project was a complete, even preposterous, failure" (Wernick, 2001, p. 5). At a certain point, "[i]t became common for writers in the British periodical press to ridicule Comte as an egotistical, eccentric, tedious, humourless, and artless writer" (Dixon, 2008, p. 43). Harriet Martineau, who translated Comte's *Course on Positive Philosophy* into English, claimed that Comte has "lost sight of his own positive principles", and refused the task of translating the *System of Positive Polity*, despite Comte's request (Dixon, 2008, p. 60). And while Mill admired the early works of Comte¸ he dismissed the *System* as the work of "a morally-intoxicated man" who did not understand the role or morality and the limitations of human beings[79] (Mill, 1865/1969, p. 336). It is in the context of Comte's audacious and extravagant *System of Positive Polity* that "altruism" was born.

In this section, I have presented an overview of Comte's philosophy, presenting the context in which the concept of altruism was created. This section illustrates how altruism was never part of a rigorous scientifically based theory. It was part of a project that would have been inconceivable today. The theoretical context that brought altruism to the intellectual world might help to explain the lack of a coherent unified view of what altruism is. In the next section, I present Comte's "cerebral theory", which explains his views on the relationship between altruism and egoism, allowing us to have a better understanding of Comte's altruism. Comte

[79] The changes we can see from the *Course* to the *System* were strongly influenced (and perhaps explained to some degree) by Comte's connection with Clotilde de Vaux, to whom he had an "ardent, chaste, and ultimately mystical attachment" (Dixon, 2008, p. 42). As Mill (1865/1969) explains, Comte thought the deep veneration he felt towards Clotilde was an example of "the sympathetic culture proper for all human beings" (p. 331). Comte thought that women represent the affective element in human nature, and the love and devotion directed at women was a preparation for the love of humanity (Comte, 1851/1875, Vol. 1, p. 211). "Saint Clotilde", as Comte called her (1851/1875, Vol. 1, p. xliv), died in his presence — a fact mentioned in a long dedicatory at the outset of the *System*.

believed that his cerebral theory offered the biological basis for his understanding of altruistic instincts, which was an important idea underlying his whole project.

7.4 Comte's Cerebral Theory and the Altruistic Instincts

"The innateness of the benevolent instincts and the earth's motion are the most important results of modern science" (Comte, 1854/1877, Vol. 4, p. 18). This passage illustrates Comte's enthusiasm for the idea of having a scientific basis for our altruistic instincts. The innateness of such benevolent instincts is based on Comte's understanding of the human mind. Such an understanding is based on an ambitious theory of the brain: Comte aims to define the number of brain parts, locate them in the brain, and explain what mental functions they are responsible for. How could Comte do that? This will be discussed in this section.

While writing long volumes in his modest office in mid-19th century Paris, Comte did not envisage anything like modern neuroscience. He intended to have a scientific basis for his theory, but he was subjected to the scientific and technological limitations of his time. A popular scientific approach to the human mind at the time was the now-infamous pseudoscience of *phrenology* (see McVeigh, 2020). Comte's cerebral theory is based on the work of phrenologists, especially the founder of phrenology, Franz Joseph Gall (Comte, 1851/1875, Vol. 1, p. 541).

Despite its negative reputation today, phrenology contributed to the progress of what we today call neuroscience. Phrenologists were pioneers in the study of the brain, defending views controversial at the time, such as the view that the brain is the seat of cognition, departing from the view that cognition was based on a "mental substance" (McVeigh, 2020, p. 332). They also

proposed that the brain is composed of specialized areas responsible for intellectual and affective faculties, a view that is now commonly accepted (Wickens, 2014, p. 135).

The negative aspect of phrenology lies in what we might call its divinatory practices. Phrenologists thought that the brain organs could shape the skull, depending on their size. So, through the measurement of the bumps in one's skull, they thought that they could make predictions about one's character (Wickens, 2014, p. 141). This practice of attributing psychological traits based on people's morphology not only was lacking empirical support, but also ended up giving support for discriminatory practices, which have garnered phrenology its bad reputation.

Although Comte relied on the work of phrenologists, he had some reservations. He complained that "mere anatomical study would never have led to the discovery of the plurality of organs", and "[a]ll attempts to determine it [the location of organs] by direct observation end in nothing but interminable disputes" (Comte, 1851/1875, Vol. 1, p. 546). Comte criticized phrenologists for analyzing the brain in isolation from the body and for their "speculative and superficial efforts at classification" (McVeigh, 2020, p. 334). But regardless of his criticism and some of his important suggestions[80], it is difficult to see how Comte's own method is any better. Comte (1851/1875, Vol. 1) believed that the appropriate method to understand the cerebral organs was studying their functions (p. 546). If this subjective method of understanding the mind's functions was properly conducted, he thought, it would enable us to discover "the number and the locality of cerebral organs" (Comte, 1851/1875, Vol. 1, p. 548).

[80] E.g., Comte suggested that the study of the brain should have comparative studies between human and non-human brains, as well as a deeper attention to pathologies — two methodologies responsible for much progress in neuroscience years later (see McVeigh, 2020, p. 334).

Regardless of its problems, it is important for us to understand Comte's view of the brain, as it tells us much about how he conceived altruism. The starting point of Comte's theory is Gall's idea that the brain is "an assemblage of several organs" (Comte, 1851/1875, Vol. 1, p. 544). Each of the brain organs is responsible for a certain relatively independent psychological function. Comte proposes that there are *eighteen* brain organs in total, which are divided into three categories, the *heart*, the *intellect*, and the *character* (Comte, 1851/1875, Vol. 1, p. 553). Comte claims that the heart is constituted of *ten* different organs, also called *affective* organs. Both Comte and Gall believed that these affective organs are predominant over the intellectual organs, and this physiological view was crucial for Comte's positivist ideal of the submission of the intellect to the heart. Comte dedicated much attention to the affective organs, and it is precisely in classifying these ten affective organs that he uses the terms "egoism" and "altruism". So, my attention in this section will be focused on Comte's views of the affective organs.

There is a twofold rule guiding Comte in his inferential process to establishing the locality of the affective organs. First, he establishes that the closer an organ is to the back of the brain, the *stronger* this organ is. This scale of strength is correlated with another scale, which goes from the more personal or *egoistic* to the more social or *altruistic*[81] organs (Comte, 1851/1875, Vol. 1, p. 558). Putting the two scales together, we have a twofold rule which states that the brain organs (and their related mental functions) "are higher in quality and inferior in force as we proceed from behind forwards" (Comte, 1851/1875, Vol. 1, p. 560). As we proceed from the back to the front, the *weaker* and the more *altruistic* the organs are. Thus, egoistic

[81] The term "altruism" was introduced only in the end of the first volume of the *System*. Before talking about "altruistic instincts" and "altruistic feelings", Comte used terms such as sympathetic instincts and social feelings, using all these terms interchangeably (see Comte, 1851/1875, Vol. 1, pp. 565, 566).

instincts are stronger than altruistic instincts, which Comte considers to be a consequence of the fact that egoistic instincts are more important for maintaining one's life. But while egoistic instincts need to be strong, they do not need much cognitive sophistication, since "[e]goism has no need of intelligence to perceive the object of desire; it has but to discover the modes of satisfying it" (Comte, 1851/1875, Vol. 1, p. 560). Comte locates them at the back of the brain, closer to the "motor apparatus" and the "vegetative viscera", away from the intellectual faculties located in the front of the brain (see McVeigh, 2020, p. 333). For altruistic instincts, we have the opposite: they are weaker and more cognitively demanding.

In the paragraph above, altruism and egoism are terms used to represent a broad category, which includes different mental functions. Comte divides the heart into ten different organs, and each of these organs is responsible for one function or instinct. The five first functions are the sexual, the nutritive, the maternal, the military, and the industrial (Comte, 1851/1875, Vol. 1, pp. 560-563). Comte considers these five to be *purely egoistic.* After them, Comte includes pride and vanity as intermediate between egoism and altruism, for although they aim for the good of the individual, they still depend on socialization (Comte, 1851/1875, Vol. 1, p. 564). The last three instincts are the ones that Comte considers to be *purely altruistic*. These are *attachment*, *veneration*, and *universal love* (Comte, 1851/1875, Vol. 1, p. 566). These three are sorted in order of excellence, so that universal love, or "humanity", as Comte also calls it, is the most perfect form of altruism. Different from attachment and veneration, which may be directed to *one* individual, universal love is always directed to a *group* of individuals. Minimally, universal love is directed towards a tribe or a community (Comte, 1851/1875, Vol. 1, p. 568), and in its most perfect form, it is directed towards humanity itself.

A reader already familiar with the contemporary literature on psychological altruism might be surprised by Comte's classification of the maternal instinct as egoistic and not altruistic. After all, much of the debate on altruistic motivation today (e.g., the evolutionary argument) is based precisely on the maternal desire to help offspring. Nonetheless, in Comte (1851/1875, Vol. 1), maternal instincts are even more egoistic than pride and vanity (p. 564). Comte explains that the maternal instinct is an "egoism of an indirect kind, which, without ceasing to be personal, yet springs from the relations of the individual to his fellow-beings" (Comte, 1851/1875, Vol. 1, p. 560). In *The Catechism of Positive Religion*, Comte (1852/2009) explains that the reason why maternal instinct might seem altruistic is that it is often *accompanied* by altruism, but this is not necessarily the case[82] (pp. 181-183).

The paragraph above raises a question about whether Comte's altruism is represented in the contemporary account of altruistic motivation. Some authors, such as Batson, believe that it is. Batson (1991) claims that the contemporary account of altruistic motivation aims to translate Comte's account into a modern language, *keeping its basic meaning* (p. 5). This became a common assumption in the literature, and authors discussing altruism take for granted that Comte defined altruism in terms of the standard account of altruism, offering Batson as a reference for such a claim (e.g., Coulter et al., 2007). But this is an oversimplification of Comte. In Comte, "altruism" refers to different things. In one sense, altruism refers to the love of humanity. This is manifest in the last of the ten affective functions of the mind, which Comte calls "universal love", and is the altruism on which Comte bases the moral code of positivism. However, as I

[82] Comte's view of maternal instincts as egoistic also reflects the views of other authors of his time. Gall also placed the maternal instincts at the back of the brain, being the second brain region, only preceded by the sexual instinct (Wickens, 2014, p. 142).

said, Comte also calls two other functions “altruistic”, namely, attachment and veneration. These differences should be taken into account before accepting or rejecting Batson’s claim.

I propose that we can conceive *two senses* of altruism in Comte. There is altruism in a *loose sense*, which encompasses the manifestations of all three altruistic instincts (attachment, veneration, and universal love), and altruism in a *strict sense*, which encompasses only the manifestations of the universal love directed at humanity. I will argue that neither the loose nor the strict senses of altruism can be properly represented by the contemporary account of altruistic motivation. Consider, first, the strict sense.

The first point to consider is that Comte’s altruism as the love of humanity (strict sense) is a morally positive notion. Love of humanity, if practiced by all, would make everyone better off. In Comte’s (1851/1875, Vol. 1) words, such altruism “can be shared by all simultaneously, not merely without antagonism, but with an increase of pleasure resulting from such community of feeling” (p. 565). These altruistic instincts admitted a “perfectly universal and boundless expansion” (Comte, 1852/1875, Vol. 2, p. 123). This is a very important feature of this kind of altruism, as it is certainly not true for the standard account of altruistic motivation, which can motivate all sorts of antisocial behaviors, as long as they aim to benefit someone other than the agent. Comte’s “to live for others” is not merely about “benefiting someone other than the individual”, but about benefiting society or humanity in general. Like Mozi, Shaftesbury, and Hutcheson, Comte was concerned with the *scope of altruism*, which is something neglected in the contemporary account. In the next chapter, I discuss the scope of altruism in more detail.

Another way of illustrating the divergence between Comte’s altruism as love for humanity (strict sense) and the standard account of altruistic motivation is the fact that Comte accepts that such altruism can be motivated by egoistic impulses. In the second volume of the

System of Positive Polity (1852/1875, Vol. 2), Comte discusses how egoistic instincts can lead to altruistic instincts, saying that self-preservation "is frequently able to awaken much attachment, and even veneration" (p. 138). The same goes for sexual and maternal instincts. Comte explains that the egoistic and the altruistic regions of the brain work together in such a way that the egoistic functions compensate for the natural feebleness of the second.

> The sympathetic instincts are rarely sufficiently strong to produce directly any very decisive action. Thus the motive of nearly every sustained course of activity arises almost invariably from some personal instinct. Even where the object is strictly social, it is impossible wholly to avoid this fatal consequence of our cerebral imperfection. (Comte, 1852/1875, Vol. 2, p. 139)

This passage suggests that Comte's treatment of altruism cannot be easily represented by the idea of an ultimate desire to benefit others. For him, an ultimate egoistic basis does not seem to preclude our feelings from being genuinely altruistic. Comte (1852/1875, Vol. 2) says that the "victory of Altruism over Egoism" could be achieved "by the indirect assistance even of the most purely personal of the instincts" (p. 139). One of the reasons making Comte (1852/1875, Vol. 2) comfortable with such an egoistic basis for altruism was his belief that, once the egoistic instincts bring us to altruistic feelings, the "irresistible charm" of altruism would allow it to take control of our conduct (p. 139).

What can we say about the *loose sense* of altruism in Comte? It is often assumed that Comte's altruism is a motivational account of altruism (Batson, 1991, p. 5; Batson, 2018, p. 22; Dixon, 2008, p. 3). This seems to be true for Comte's altruism in the strict sense, discussed above. But when it comes to Comte's altruism in a *loose sense*, it is not obvious that it is motivational at all. Remember that Comte's altruism in a loose sense is applied to all three altruistic instincts (attachment, veneration, and universal love). These do not seem to be

dependent on the higher-order motivations characteristic of the standard account of altruistic motivation. If this is true, then Comte's altruism in the loose sense certainly cannot be represented by the standard account of altruistic motivation.

Perhaps a good piece of evidence pointing to the conclusion that altruism in a loose sense in Comte does not involve higher-order motivational states is his view of altruism in other animals. Although Comte thought that altruistic instincts reach their most perfect state in humans, he accepted that they are present in many non-human animals. He thought that, while some species are almost exclusively concerned with themselves, others are more altruistic. For example, attachment is observed in monogamous species, and veneration is observed in dogs towards their owners (Comte, 1851/1875, Vol. 1, p. 567). Comte thought that dogs and stags can "live for others" to a certain degree (Dixon, 2008, 51). In this sense, altruism (in the loose sense) is a trait widespread in nature, and does not seem to require a higher-order motivation.

Even if we restrict Comte's altruism in the loose sense to motivations, it would be odd to restrict these motivations to the higher-order kind of motivation present in the contemporary account of altruistic motivation. Comte's altruism was more inclusive, being manifested in different animals. There are no good reasons for limiting altruism in the loose sense to the specific kind of motivation we today call altruistic motivation.

My conclusion, therefore, is that neither Comte's altruism in a strict sense nor Comte's altruism in a loose sense can be represented by the standard account of altruistic motivation. As Comte's altruism is the original account of altruism, such a divergence has important implications for our discussion about the proper definition of altruism, which will be addressed in the next chapter.

In this section, I discussed Comte's theory of the brain, explaining the context in which altruism was conceived. Such a theory has many flaws, partly due to Comte's ambitious project, and partly due to the scientific limitations of his time. Comte (1851/1875, Vol. 1) himself was aware of these limitations when he said that he left to his successors "the task of ultimately employing the objective method" in the study of the mind "as soon as the time for it shall have arrived" (p. 547). More importantly for us, however, is how Comte's cerebral theory has shown the particularities of Comte's account of altruism. As this section shows, altruism as the love of humanity (the strict sense of altruism), is at odds with the contemporary account of altruistic motivation. In the next section, I will discuss how altruism fits into the evolutionary accounts of the 19th century, showing how the evolutionary account of altruism was entwined with altruism since its origins.

7.5 Altruism Meets Evolutionary Theory

For Comte (1851/1875, Vol. 1), the goal of morality was "subordinating as far as possible the personal to the social instincts, by referring all to Humanity" (p. 559). He celebrated that "modern biology demonstrated that altruistic propensities were actually innate to humanity" (Dixon, 2008, p. 43). However, as explained in the previous section, even if altruism is innate, it is still *weaker* than egoism. So, a crucial question is: how can we rely on our altruistic propensities as the basis for society and morality? How can Comte's ideal of "to live for others" become a reality if the altruistic impulses are weaker than the egoistic impulses? To understand Comte's argument for how altruistic instincts may prevail, we need to address his pre-Darwinist understanding of evolution.

One of the fundamental laws in animal life, for Comte (1851/1875, Vol. 1), is the "law of exercise" (p. 211). In this Lamarckian law, the use of organs (including that of the brain) could be developed or atrophied based on our use or disuse of them, and such a change could be *inherited* by the following generations (see Dixon, 2008, p. 52). It is based on this law that Comte thought we could expect the brain organs responsible for our altruistic instincts to overcome egoistic instincts: altruism could be strengthened by regular exercise. The gradual progress of humanity, more easily providing the basic needs for a good life, would gradually reduce the *need* for egoism, which at the current stage is so necessary.

In Comte's evolutionary view, as time passes, altruistic instincts would become more pervasive. Thus, through the law of exercise, future generations would become more altruistic (Comte, 1852/1875, Vol. 2, p. 127). Such an evolutionary process would allow humanity to reach a stage in which morality would no longer require theological beliefs or self-interest as its foundation. The evolutionary approach to altruism as the foundation of a naturalistic ethical account was shared by other authors. Herbert Spencer, one of the most important intellectuals of the 19th century, developed a similar approach.

A common accusation leveled against the British evolutionary theorists of the 19th century, such as Spencer and Darwin, is that these authors' evolutionary views were perpetuating the Victorian prejudices of their time[83] (Todes, 1989). Ingold (1986/2016) claims that both Darwin and Spencer follow "a well-established tradition of liberal social philosophy, according to which society is rationally constituted as an instrumental adjunct to the satisfaction of

[83] The use of metaphors such as "survival of the fittest", which was coined by Spencer and later used by Darwin (Weinstein, 2019), and "struggle for existence" arguably illustrate the assimilation of the British spirit of the time into evolutionary theory. For a criticism of the view of British authors as perpetuating Victorian prejudices, see Dixon (2008).

extrasocial and purely hedonistic end" (p. 186). The idea of a relentless competitive nature was at home with the British *ethos* of the time, which was marked by the ascension of industrialization and the widespread influence of libertarian ideals (see Ingold, 1986/2016, pp. 201-202). Authors such as Kropotkin (1902/2021) thought that the British overly competitive conception of nature, especially illustrated in the work of T. H. Huxley, also known as "Darwin's Bulldog", was a biased view that did not find support in evidence from evolutionary biology[84] (see Harman, 2014).

However, despite the views above, Spencer and Darwin were much concerned with our altruistic dispositions and their moral implications. Like Comte, Spencer was "sensitive to the problem of finding a new and natural ethic to replace the moral code which had been associated with the traditional faith" (Durant, 1962, p. 358). Spencer thought that such a problem would not be overcome through a top-down imposition of artificial norms. Instead, Spencer argued that the *altruistic* dispositions of humanity can evolve and overcome their egoistic counterparts. Humanity — which Spencer considered to be a "superorganism" (Ingold, 1986/2016, p. 185) — has to slowly *evolve* into a higher social state. With the progress of society, our sympathetic nature could become pervasive, not depending on moral principles — we would pursue altruistic behaviors as we pursue any other pleasure. "[I]ndustry and peace will develop altruism to the point where it will balance egoism" (Durant, 1962, p. 363). Spencer, again like Comte, postulated that humanity passes through distinct "stages" of evolution[85] (Weinstein, 2019).

---

[84] In a passage that is often used to illustrate his harsh view of evolution, Huxley (1888/1902) describes the animal world as a "gladiator's show", in which "[t]he creatures are fairly well treated, and set to fight — whereby the strongest, the swiftest, and the cunningest live to fight another day" (pp. 199-200).

[85] For Spencer, there are *four* stages rather than three as Comte argued. Nevertheless, it is clear the similarity between the two authors. As Dixon (2008) comments, "Comte and Spencer both wrote turgid, multi-volume, jargon-packed philosophical syntheses. Both proposed their own

Darwin (1871/2009) shared Comte's and Spencer's evolutionary views of ethics (pp. 73, 97). In his *Descent of Man* (1871/2009), Darwin joins these authors in their optimism about the evolution of our sympathetic instincts, claiming that these instincts may become fixed through inheritance (p. 104). Although Darwin never used the term "altruism" in his writings, many have considered him to be "a theorist of altruism and selfishness" (Dixon, 2008, p. 5). Like Comte and Spencer, Darwin was "trying to work out an understanding of morality which was grounded in innate instincts of 'love' and 'sympathy'" (Dixon, 2008, p. 132).

Despite their interest in finding a non-egoistic basis for morality, we should not commit the mistake of claiming that Spencer and Darwin were proponents of psychological altruism. Although Darwin is concerned with rejecting an account of morality based on selfishness, the account of selfishness that he refers to is similar to the deliberate egoism discussed in *Section 7.2* (see Darwin, 1871/2009, p. 86). For Darwin (1871/2009), "the most noble part of our nature" is not selfish, "unless indeed the satisfaction which every animal feels when it follows its proper instincts… be called selfish" (pp. 98-99). The egoistic internal rewards, which occupy the contemporary debate, were not the sort of thing that would make something selfish for Darwin.

In the case of Spencer, it is even clearer that his altruism should not be equated with the contemporary account of altruistic motivation. Altruism, in Spencer, has a hedonistic foundation (see Ingold, 1986/2016, p. 230). His explanation for how altruism can evolve and why one should be altruistic is based on one's pursuit of pleasure or happiness (see Spencer, 1862/1883,

---

classification of the sciences, both held the sciences of sociology and ethics to be the most important of all, and both were closely associated with 'altruism'" (p. 195). Spencer, however, claimed that Comte did not have a substantive influence in his work (see Dixon, 2008, p. 203). While some authors accept Spencer's claims about not being influenced by Comte (Eisen, 1967), others suggest that Comte's influence could have been indirect or even unconscious to Spencer (Durant, 1962, p. 363; Dixon, 2008, p. 205).

p. 193). Our welfare depends upon the welfare of others, so, for egoistic reasons, we should care about their welfare (Spencer, 1862/1883, pp. 205-211). Similarly, Darwin also talks about the motivation to care for others in terms of the pleasure that these actions cause us (see Darwin, 1871/2009, p. 72).

But can we make sense of Spencer's altruism in terms of altruistic motivation as we understand it today? In early works, Spencer regarded altruism as a *sentiment* (Dixon, 2008, p. 197). So, this could be interpreted as a motivational account of altruism. However, in his *Data of Ethics* (1862/1883), Spencer interprets altruism differently. In this work, Spencer adopts an evolutionary account of altruism, rather than a motivational account. He claimed that an action is altruistic if it causes a cost to the performer and a benefit to the receiver. "Whatever action, unconscious or conscious, involves expenditure of individual life to the end of increasing life in other individuals, is unquestionably altruistic" (Spencer, 1862/1883, p. 201). These actions do not require consciousness nor any mental representation (Spencer, 1862/1883, p. 201). This is very similar to the contemporary account of evolutionary altruism, but instead of the contemporary "fitness" to refer to the unit being decreased in the agent and increased in the recipient, Spencer uses the term "life" or "bodily substance" (see Spencer, 1862/1883, p. 203; see Ingold, 1986/2016, p. 229). Dixon (2008) claims that the evolutionary account of altruism "first emerged through the influence of Herbert Spencer's writings" (p. 3).

The change from a motivational to an evolutionary account created much confusion for his readers[86]. But it illustrates that the fragmentation of altruism into different accounts was

[86] Later on, in the 1890s, Spencer distanced himself from the term "altruism" altogether. This change, Dixon (2008) argues, was ultimately motivated by strategic and political reasons, since by then altruism was a word associated with socialism and Christianity, as well as Comte's positivism (pp. 209-210). Spencer did not want to be associated with any of these doctrines.

already in place in the 19th century. It seems that the importance of altruism for ethics, on the one hand, and its evolutionary character, on the other hand, creates a tension between a conception of altruism as a motivation and a behavior. This tension is still present today. The division between behavioral and motivational accounts of altruism might not be the best way of dealing with the complex nature of altruism. In Chapter 9, I propose a new account of altruism, virtue altruism, which brings together motivation and behavior, thus avoiding this tension.

In this chapter, I discussed different authors who, in one way or another, have emphasized the non-egoistic motivations present in human beings. From this survey of ideas, we can see that much of what these authors argued would be missed if we reduced the motivations they discuss to what we have today taken to be altruistic motivation. Many issues are overlooked if we make this reduction: what is the proper *scope of altruism*? Who are the "others" that altruists care about? Should we include only certain kinds of people in the scope of our concern? Does our altruistic motivation need to be morally praiseworthy? Can we have a proper account of altruism that is completely devoid of moral value? The next two chapters discuss these aspects that are neglected in the standard account.

Chapter 8

Cost, Scope, and Action: The Forgotten Traits of Altruism

8.1 The Special Status Hypothesis

Part I of this thesis presented the many faces of altruism in the contemporary literature and discussed the framework of altruistic motivation. Part II discussed the arguments for psychological altruism in the contemporary scientific literature. Part III looked back to the history of altruism and egoism, discussing how the contemporary debate diverges from the historical views addressed. Now, in Part IV, I address the non-technical, ordinary use of the term "altruism" and its often-neglected normative character. As I mentioned in the introduction, I will use the term "ordinary altruism" to refer to the meaning of altruism as it is conceived in everyday, ordinary language. In this chapter, I discuss the relationship between the standard account of altruistic motivation and ordinary altruism, arguing that the former fails in representing the latter properly. In the next chapter, I address the normative character of altruism and propose some alternatives to rethink altruistic motivation, moving away from the limiting notion of ultimate desires.

Since the definition of altruistic motivation in terms of ultimate desires is the standardly accepted view in the literature, authors often adopt it without giving further justification for why they do so. However, sometimes authors try to justify their adoption of this particular account of altruism. A common strategy in doing so is to appeal to the reader's *intuition* about what altruism is. Stories and thought experiments are presented as a means to show that the standard account is coherent with what the readers would identify as altruistic. The following passage is an example

of this. Sober and Wilson (1998) discuss why the standard account of altruistic motivation should be preferred over a broader account of altruism.

> [C]onsider a heroin addict whose every action is ultimately aimed at securing the pleasant states of consciousness that the drug produces…. Let us place this person in an environment in which the only way to get the drug is by helping people... An addict who helps others only because the effect of helping is a drug-induced euphoria is not thereby an altruist. The same point applies to people in the real world who do not take heroin; if they are "hooked" on helping because of the pleasure that helping affords and the pain it allows them to avoid, their actions do not make them altruists. (Sober & Wilson, 1998, p. 230)

In this passage, Sober and Wilson are not merely explaining what follows once we accept the standard account of altruistic motivation. Instead, they are appealing to the reader's intuitions about what should be called "altruistic", claiming that the standard account of altruistic motivation is more at home with these intuitions than the competing alternative accounts. Appeals to readers' intuitions like this are common in the literature. The entry on altruism in the authoritative Stanford Encyclopedia of Philosophy (Kraut, 2020), for example, makes claims about what people would "normally" consider being altruistic and about what would be "odd" to call altruistic.

The appeals to the readers' intuitions about altruism suggest that these authors take the standard account of altruistic motivation to be the best representation of commonly held, ordinary intuitions about altruism. Although this is usually an implicit assumption, in some cases authors make a more explicit defense of this idea. In an early paper, Elliott Sober (1988) discusses what he calls "vernacular altruism", which aims to represent altruism as it is used in everyday language. He starts by claiming that the key feature of vernacular altruism is that it is motivational. Whatever altruism is, Sober (1988) says, it "has to do with motives" (p. 76). He then specifies that such a motivation consists of non-instrumental other-directed desire (Sober,

1988, p. 77). This is basically the standard account of altruistic motivation, which Sober calls "psychological altruism" in later works (e.g., Sober, 2013).

Sober's (1988) reduction of vernacular altruism to ultimate altruistic desires echoes Daniel Batson's views. Batson (1991) considers the account of altruism he uses to be representative of both its original Comtean version and its meaning in everyday language (p. 5). Batson (1991) dismisses competing accounts of altruism, calling them "pseudo-altruistic" (p. 43). This view that the standard account represents altruism as it is in general remains present in his later work (see Batson, 2011, 2018).

At this point, a legitimate concern that a reader might have is that, by looking at the scientific and historical dimensions of altruistic motivation, I am ignoring this special link between the standard account and ordinary altruism. What if, despite all the problems I raised, the standard account of altruistic motivation is successful in reflecting the general, non-technical use of altruism, capturing our commonsense intuitions about what altruism means? If this is the case, then the standard account of altruistic motivation deserves a special status when compared to other technical accounts of altruism. I will call this the *special status hypothesis*: the standard account of altruistic motivation is not one technical account among others, but the best representation of ordinary altruism, thus deserving a special status when compared to these other technical accounts.

If the special status hypothesis is true, there might still be good reasons to hold the standard account of altruistic motivation. Even if I am correct regarding the criticism I raised in previous chapters, the view of altruistic motivation as an ultimate desire to increase the welfare of others would still be legitimate if it reflects ordinary altruism. In this chapter, I respond to this problem. I offer reasons to *reject* the special status hypothesis, claiming that the standard account

of altruistic motivation fails in representing ordinary altruism. The next section explains the challenges in proposing a definition that accounts for ordinary altruism and presents the argumentative strategy that I will adopt to reject the special status hypothesis.

8.2 A Strategy Against the Special Status Hypothesis

Providing evidence against the special status hypothesis is not an easy task. If we aim to reject the claim that the standard account of altruistic motivation is the best representation of our non-technical, ordinary use of "altruism", we need to identify the criterion used in ordinary altruism. This, however, is hard. One should be careful in trusting one's own intuition about what this ordinary use is. As Sober (2013) warns us, "[p]hilosophers need to be careful not to confuse common sense with what they themselves happen to find obvious" (p. 140).

The ordinary use of some categories can be illustrated by the scientific definitions of the term. However, in the case of altruism, we do not have a single, unequivocal definition. There are many technical accounts of altruism, with quite divergent features. So, although they can help us to indicate a broad cluster of features associated with altruism, they do not help in determining which account is the best representation of altruism.

Looking at dictionary entries does not help much either. The Merriam-Webster dictionary (n.d.) defines altruism as an "unselfish regard for or devotion to the welfare of others" and as a "behavior" that benefit others. In the Cambridge Dictionary (n.d.), altruism is defined as a "willingness to do things that bring advantages to others, even if it results in disadvantage for yourself" and as "the attitude of caring about others and doing acts that help them although you do not get anything by doing those acts". So, the simple willingness to help can count as

altruistic in some cases, while a costly behavior that helps others may count as altruistic in other contexts. Rather than pointing to a clear, specific definition, these entries indicate that the word "altruism" refers to a cluster of features, involving different behaviors and motivations.

Another issue making ordinary altruism hard to grasp is that this word is simply not used very often. This, as Schefczyk and Peacock (2010) argue, makes its meaning not so well-integrated into everyday language as that of other words. Thus, it is hard to know when a given definition of altruism is in reflective equilibrium with common sense (Schefczyk & Peacock, 2010, p. 166). Thus, considering these challenges, there are reasons to be skeptical of a clear definition that reflects ordinary altruism.

Considering the challenges mentioned above, the strategy I will adopt in this chapter is to investigate prototypical cases of altruism and use this as a means to test the special status hypothesis. First, I will discuss different ways of conceiving categories. After this, I will explain why I believe that we are better off conceptualizing altruism as a category without clear-cut boundaries. finally, I will explain how the notion of prototypical cases can be used to reject the special status hypothesis.

Research on lexical semantics has proposed many different ways of understanding categorization. In the traditional approach, "concepts are clearly defined by a conjunction of singly necessary and jointly sufficient attributes" (Hampton, 1999, p. 178). This way of thinking about categories is widespread not only in science, but in the Western tradition as a whole (Armstrong et al., 1983, p. 266). Of course, establishing necessary and sufficient conditions has many advantages. Categories with clear boundaries allow us to determine, with precision, whether something fits into a category or not. So, it is not surprising that all the technical accounts of altruism are defined in terms of clear necessary and sufficient conditions.

However, philosophers and scientists have disputed the traditional approach described above. They claim that not every concept can be determined with clear necessary and sufficient conditions. This view became influential through the work of Ludwig Wittgenstein (1953/1968). He famously proposed that, in some cases, the members of a category do not share common essential features. Instead, they can share features with some members of the category, and these members can share features with others, and so on. For example, in a category composed of A, B, C, and D, individual A can share features with B, while B shares features with C, and C shares features with D. These overlapping sets of features can constitute a category, even though no property is shared by all members. Members are related through *family resemblance* (see Wittgenstein, 1953/1968, p. 32). Individual A might not share any feature with individual D, but they can still belong to the same category. This idea offers a way of thinking about a category without establishing necessary and sufficient conditions[87].

The work of the psycholinguist Eleanor Rosch (1973) has offered empirical basis for the view of categories without clear boundaries. Rosch's work has shown that people understand categories in terms of their best examples rather than necessary and sufficient conditions. Rosch (1978) uses the term "prototype" to represent the "clearest cases of category membership" (p. 36). As the philosopher Ian Hacking (1995) explains, "[e]ach class has best examples" and other examples that "radiate away" from these examples (p. 23). For example, a robin is considered to

[87] Categories without clear boundaries also became an influential idea in biology. Richard Boyd's homeostatic property cluster (HPC) offers another philosophical approach that allows us to depart from necessary and sufficient conditions. Boyd (1988/1995) proposes that some natural kinds consist of properties that frequently co-occur in nature. These properties are clustered by certain mechanisms, which make the presence of one of these properties increase the likelihood of the presence of the others. Natural kinds, thus, can refer to these homeostatic clusters, thus not having any particular property that is either necessary or sufficient. Boyd's homeostatic property cluster offers an interesting way of conceptualizing categories without the need to establish clear-cut boundaries.

be a prototypical bird, while ostriches and pelicans are considered marginal cases (Armstrong et al., 1983, p. 269).

Rosch (1978) claims that identifying prototypical cases is a fundamental aspect of our normal way of thinking about categories. Importantly, her research supports the view that people agree about how good an example is to represent a concept e*ven when they disagree regarding the limits of the concept* (Rosch, 1978, p. 36). This is crucial for the investigation of ordinary altruism. The idea of prototypical cases allows us to have access to a category even when we do not know its necessary and sufficient conditions, and this is precisely the problem we have. On the one hand, we have a broad cluster of properties associated with altruism. On the other hand, we still have intuitions about what is more or less altruistic. I claim that, if we conceptualize altruism as a cluster-type category, with prototypical cases in the center and other cases that radiate away, then we can make sense of ordinary altruism — this is the approach I will propose in this chapter.

There is a caveat that needs to be mentioned. The empirical work of researchers like Rosch shows that individuals can identify prototypical cases of categories. Rosch (1978) claims that "[m]ost, if not all, categories do not have clear-cut boundaries" (p. 35). However, as Armstrong et al. (1983) convincingly show, people can identify prototypical cases even in well-defined categories. More recently, Lupyan (2013) has shown that even for categories such as "triangle", there are cases consistently identified as best examples. But this does not mean that these concepts are defined on the basis of these prototypical examples. The discussion of the prototypicality effects, as a "scalar goodness-of-example", should be distinguished from the discussion on the nature of categorization (see Lakoff, 2007, p. 132). Armstrong et al. (1983)

propose that we can explain the empirical findings of prototypicality as an "organization of 'exemplariness'" rather than an "organization of 'class membership'" (p. 292).

Some phenomena in scientific research are traditionally represented through prototypical cases rather than necessary and sufficient conditions[88]. However, my argument in this chapter does not make assumptions about the nature of categorization in general. All I need, here, is to assume two more modest claims: (1) that *some* categories do not have clear-cut boundaries, being better explained in terms of their prototypical cases; and (2) that ordinary altruism is better explained as a category without clear-cut boundaries, with prototypical cases in its center. The work of philosophers like Wittgenstein and scientists like Rosch supports (1). The broadness of dictionary entries and the existence of a large range of diverging technical definitions of altruism in the scientific literature support (2).

The method I will use to investigate ordinary altruism is to analyze three famous cases traditionally regarded as illustrative of altruism and to take them as prototypical cases of altruism. In *Section 8.3*, I discuss the case of Wesley Autrey, who saved the life of a stranger who fell unto the subway tracks. In *Section 8.4*, I discuss the rescuers in WW2, who risked their lives to save Jews and other persecuted minorities. Finally, in *Section 8.5*, I discuss the famous parable of the Good Samaritan.

I will use the prototypical cases mentioned above to argue against the special status hypothesis. To see how the argument will work, consider the following example. Imagine a definition that aims to represent the ordinary use of the word "bird". If this definition

[88] For example, as Hacking (1995) explains, mental disorders are constituted by clusters of symptoms rather than necessary and sufficient properties. Hacking (1995) claims that thinking in terms of a prototype is something implicit in psychiatry, where the description of clear examples is a common way of characterizing mental disorders (p. 24).

consistently excluded prototypical cases like robins and consistently included peripheral cases like ostriches, then we could conclude that this definition fails in representing the ordinary use of the word "bird". Analogously, if the special status hypothesis is correct, then the standard account of altruistic motivation should do a good job of representing the prototypical cases of altruism. However, as I will show, it fails in doing so. The standard account not only fails in capturing prototypical cases of altruism, but it also captures cases that would not be regarded as altruistic in the ordinary use of "altruism".

The literature on prototypical theories is vast and complex. There are many ways of understanding prototypes. For example, we can conceive prototypes as examples more closely related to the goals of a category, as examples associated with the central tendency of a category, or as examples associated with the most common occurrences of a category (Barsalou, 1985). Also, there is debate about whether we should focus on prototypical features or prototypical examples (see Hampton, 1999, p. 178). Debating the particularities of this literature goes beyond the scope of my project.

For the purposes of my argument, I will consider that there are prototypical *examples* of altruism and that there are prototypical *features* of altruism. The cases I will discuss in the next sections are prototypical cases, and they illustrate prototypical features. Each of the examples presented aims to illustrate a different feature of altruism that is neglected by the standard account of altruistic motivation. *Section 8.3* highlights the costs of altruism, *Section 8.4* the scope of altruism, and *Section 8.5* the need for altruistic actions. I will consider the three cases as prototypical cases of altruism.

In the previous section, I introduced the special status hypothesis, claiming that it is explicitly adopted by some authors and implicitly adopted by many others. In this section, I

discussed the challenges in representing ordinary altruism and proposed that the analysis of prototypical cases of altruism offers a way of testing the special status hypothesis. In the following three sections, I will propose an investigation of three prototypical examples of altruism. This chapter aims to reject the special status hypothesis, that is, it claims that the standard account does not accurately represent ordinary altruism.

8.3 The Costs of Altruism

I start this section by presenting the story of a construction worker named Wesley Autrey, who became known as the "subway hero" after risking his life to save a stranger who fell onto the tracks of the subway, in New York. After realizing that he would not be able to pull the man out of the tracks in time to escape a train that was approaching, Autrey threw himself over the man, holding him down between the tracks, thus saving his life (Garson, 2015, p. 7).

I claim that the case above constitutes a prototypical case of altruism. If anything deserves to be labeled "altruistic" in everyday language, it seems that Autrey's action should be so described. The word "altruism", so rarely used, is often reserved for cases such as this. The central feature highlighted by Autrey's case, which will be the focus of discussion in this section, is that altruistic actions are *costly* to the agent. Altruists engage in actions that involve either a significant cost or a significant risk of cost. People like Wesley Autrey are praised as altruists not merely for "helping others," and much less for merely "desiring to help others," but for helping others while accepting the significant personal costs of helping.

Some authors have considered how having a cost is a common trait of altruism. Oliner and Oliner (1988), for example, claim that one of the conditions for altruism is that "it involves a

high risk or sacrifice to the actor" (p. 6). However, this is not the standardly accepted view (see Sober & Wilson, 1998, p. 246). For authors such as Batson (2011), altruism "does not imply self-sacrifice of any sort" (p. 23). Altruistic motivation is considered to depend only on whether the desire to help is ultimate. It does not matter if the altruistic actions produce only advantageous outcomes for the agent. In the standard view, costs can occur, but they are not relevant[89].

But why have psychologists and philosophers adopted an account of altruism that neglects costs and considered it to represent ordinary altruism? We can find two possible answers to this question in Batson's work. The first is that focusing on costs or "self-sacrifice" "shifts attention from the crucial question of motivation to consequences" (Batson, 2011, p. 23). The second is that "a definition based on self-sacrifice overlooks the possibility that some self-benefits increase as the costs of benefiting another increase" (Batson, 2011, p. 23). For example, a saint or a hero can incur major costs, but their rewards may also increase as the costs increase.

If we simply state that costs are a necessary condition for altruism, then Batson might be right in saying that altruism becomes something more about consequences and less about motivations. However, there are better ways to account for the costs of altruism. Schefczyk and Peacock (2010) propose a *condition of reasonable expectation of cost*, which claims that "[a]n action, Φ, is altruistic if and only if the agent reasonably expects to bear net costs from performing Φ" (p. 171). This offers a way to respond to the concerns raised by Batson.

[89] Kraut (2020), also recognizing the importance of costs, proposes the term "weak altruism" for cases where the individual merely helps others and "strong altruism" for cases where the agent knows that there will be some personal loss following from the altruistic action. I avoid this distinction, however, because the expectation of costs is only one variable among others, and it is not clear why this should be regarded as the feature in which a distinction between "weak" and "strong" should be based. Moreover, we should be cautious before making such distinctions considering that they may cause more confusion. In this particular case, other authors have already used the term "weak altruism" in a different sense (e.g., Kitcher, 1998, p. 287).

As Schefczyk and Peacock explain, we do not regard someone as less altruistic if his or her altruistic actions resulted in *unexpected* benefits for the agent. Wesley Autrey, for example, was rewarded after his altruistic deed — including a $10,000-dollar cheque given by none other than Donald Trump. But these rewards did not make Autrey's action egoistic (although they could do so if Autrey knew about them *before* jumping onto the tracks). Schefczyk and Peacock's insight is that, rather than establishing a *need for a cost*, it is preferable to establish a condition of *expectation of costs*.

After the clarification presented above, consider again the concerns raised by Batson. The first reason Batson (2011) gives for neglecting costs is that it shifts the focus from motivations to consequences. But this can be easily avoided if the condition for a cost is articulated as an *expectation* of cost rather than actual costs. It is the motivation that has to include an expectation of costs, not the actual outcome that has to be costly. If I helped someone and later discovered that, in doing so, I was risking my life, this would not make my motivation altruistic. Batson's second concern is that some self-benefits increase when costs increase. However, if these rewards are not expected by the agent, their existence in the future would constitute no objection to their status as altruists, as Autrey's case illustrates very well. If agents *know* that rewards will compensate for their costs, then they will *not* have a true expectation of costs. Therefore, the condition of expected cost properly responds to both concerns raised by Batson.

Notice that, until this point, I did not address Wesley Autrey's motivational states directly. Presumably, his action was voluntary, and his cognitive faculties were working properly. But was his desire to help *ultimate*? We do not know. If the analysis I provided in previous chapters is correct, neither Autrey himself, the reporters, nor Donald Trump can be sure about whether Autrey's ultimate desires were egoistic or altruistic. But the very fact that we do

not know whether Wesley Autrey had ultimate desires to help illustrates that ultimate desires might not be as important as proponents of the standard account believe. I will return to this in the next chapter, where I argue that ultimate desires are only one feature of altruism among others, not the one on the basis of which we should define altruism.

This section addressed the issue of the costs of altruism. The expectation of costs is a key feature of altruism, as illustrated in Wesley Autrey's example. If he helped someone without risking anything, even if moved by the same sort of motivation to help, this would hardly be a prototypical case of altruism. The expectation of cost that he courageously accepted when jumping into the tracks is what we can consider a prototypical feature of altruism. Drawing on the work of Schefczyk and Peacock (2010), I clarified that we should focus on the expectation of cost rather than the costs themselves. The standard account of altruistic motivation does not capture this feature, which indicates a failure in representing ordinary altruism. In the next section, I discuss another forgotten trait of altruism, namely, its *scope*.

8.4 The Scope of Altruism

The Second World War has unleashed humanity's selfish and evil sides whose existence we often (want to) forget. But among the horrors of the war, noble traits have also found space to flourish. One of the remarkable cases of such flourishing is represented in the actions of civilians who risked their lives to save Jews and other groups persecuted by the Nazi party. These rescuers have included under the scope of their concern people to whom they did not have a clear obligation. The rescued Jews were strangers to many of the rescuers. Why did these people help while so many others closed their eyes to the suffering of those in need? This is a question asked

by Oliner and Oliner (1988), who try to identify the stable personality traits characteristic of these rescuers. In this section, I discuss the case of rescuers as another prototypical case of altruism[90]. This new case will be the basis for us to explore the scope of altruism.

In Chapter 7, I discussed the importance of the scope of one's care. From the opposition between Confucius's and Mozi's accounts of love for others to Shaftesbury's idea that the moral sense approves the actions that aim for the good of mankind, we see a constant concern with the question regarding who should be the *beneficiaries* of our care. In contemporary literature, the issue has been explored by authors such as Kitcher (2011), who recognizes that altruism is a multidimensional notion. He comments that one of the relevant variables in altruism should be the range of individuals who can trigger the agent's altruistic response (Kitcher, 2011, p. 31). What are the limits of our altruism? Who is included inside the scope of our concern and who is left out?

In the standard account, altruistic motivation is an ultimate desire to increase the welfare of *others*. By "others," we should understand simply *anyone other than the agent*. It can include people with all sorts of relations with the agent, from close family members to complete strangers. But when we think about the rescuers mentioned before, something is missing if we conceive the scope of altruism in this way. A fundamental aspect of their altruism is *to whom* they directed their concern, and this is not reflected in the standard account.

[90] I could have used this example to discuss the costs of altruism as well. However, my preference for Wesley Autrey's case as the basis for my discussing on the costs of altruism is that, in his action, he was not putting others in risk. The cost of his altruistic action was his own. By contrast, in the case of the rescuers, they were also putting their family members in risk, often without their consent (Galston, 1993, p. 131). This makes the issue of the costs a little more complex in the case of rescuers.

Interested in the issue of the scope of altruism, Galston (1993) proposes a useful classification of altruism based on its scope. In what he calls *personal altruism*, the scope of altruism includes only family members, friends, and people closely related to us. In *communal altruism*, the scope comprehends people "like us," who share traits, who belong to the same ethnic group or religion, etc. Finally, *cosmopolitan altruism* "is directed toward the human race as a whole, and hence toward individuals to whom one has no special ties" (Galston, 1993, p. 123).

The fact that these rescuers presented a case of cosmopolitan altruism, rather than simply personal altruism, is relevant to our consideration of their example as prototypical cases of altruism. As Oliner and Oliner (1988) point out, "[w]hat makes their behavior of particular interest is… that it was undertaken on behalf of an 'outsider' minority group, marginal under normal conditions and increasingly rejected and despised" (p. 1). The authors discovered that, although some rescuers had previous bonds with the Jews they helped, at least *50 percent* of them had no pre-acquaintances with the individuals they helped, and "almost 90 percent helped at least one Jewish stranger" (Oliner & Oliner, 1988, p. 81). The crucial feature of the case of the rescuers is that the scope of their altruism included people to whom they had no particular obligations: they were risking their lives to save people from outside their family or close circle of friends. This is at least as important as their desire to help and the costs they incurred.

To illustrate further how the standard account of altruistic motivation diverges from ordinary intuitions about the scope of altruism, consider an example. Imagine a father who deeply loves his daughter. He always acts in ways that improve her welfare, and his desires to help her are ultimate. The father often acts in ways that are costly to himself. According to the standard account, both rescuers and the father represent cases of altruism — as long as they have

ultimate desires to help others. But although we could praise this father for being a good father, it seems unlikely that we would use the term "altruist" to describe him. Calling a father "altruist" because he deeply wants the best for his own daughter seems to be simply a misuse of the word. The people who closed their doors to escaping Jews presumably had loved family members. They might ultimately desire the good of these few people. But stretching the category "altruism" to the point of including these people who care only for their family members would deflate altruism of its central meaning as it is used in this context.

Before ending this section, I will raise one last concern: the standard account of altruistic motivation is not only neutral about the scope of altruism, but actually privileges personal altruism over cosmopolitan altruism. I suggest this because cosmopolitan altruism is directed to an impersonal group, while personal altruism is directed to a particular individual. Sober (2013) claims that if someone "desires the greatest good for the greatest number, the desire is impersonal… the desire's content singles out neither self nor *specific others* [emphasis added]" (p. 131). He suggests that, for this reason, the desire "is neither altruistic nor egoistic" (Sober, 2013, p. 131). In a similar vein, Sober and Wilson (1998) explain that "it is the impersonal character of moral principles that distinguishes them from the personal character of altruistic and self-interested desires" (p. 240).

Now, consider the case of rescuers. According to the standard account, if a rescuer helps a particular individual, she can be an altruist. But, following what Sober and Wilson (1998) and Sober (2013) seem to suggest, a rescuer who is protesting against the Nazi policies, or actively searching for members of persecuted minorities to provide help to them, might not be an altruist. If her desire is directed to alleviate the suffering of persecuted minorities, this desire could be too impersonal to be altruistic. Altruists who aim to reduce the suffering in the world would not be

considered altruists in this view, for their desire is too impersonal. This restriction selects against cases of cosmopolitan altruism, privileging cases directed to particular individuals. This restriction does not seem to be consistent with ordinary altruism, in which cosmopolitan altruism appears more clearly altruistic than personal altruism.

There is, of course, a *methodological* reason for focusing on altruistic motivation directed to particular individuals. Considering the difficulties in assessing ultimate altruistic desires, researchers have to focus on particularly intense or clear cases of altruistic motivation. For example, if altruism is an ultimate desire, mothers will be the best examples of altruists. That is because they are the ones we should, for evolutionary reasons, expect to have ultimate desires to benefit their offspring. But beyond this methodological reason for preferring personal altruism, Sober and Wilson (1998) and Sober (2013) seem to suggest a more direct dismissal of a more abstract, impersonal cosmopolitan altruism.

In this section, I argued that the standard account of altruistic motivation neglects a fundamental aspect of altruism, which is the property of being directed to people that are not in the agent's immediate scope of concern. The focus of the standard account is entirely directed to the question of whether the desire to help others is ultimate or not. However, in this section, I argued that the scope of altruism is at least as important as the question of whether the desire is ultimate or instrumental. Since the standard account does not distinguish between different beneficiaries, it is incapable of capturing a fundamental feature of ordinary altruism. Any account that ignores the difference between *personal* and *cosmopolitan* altruism will hardly reflect these common intuitions about altruism. Finally, I also suggested that some authors seem to privilege personal altruism over cosmopolitan altruism by requiring altruistic motivation to be

directed to particular individuals. The next section addresses another important feature of prototypical cases of altruism, namely, the altruistic actions themselves.

8.5 Altruistic Actions

As I argued in the previous chapter, even though the term "altruism" was coined only in the 19th century, the general idea of benefiting others for their own sake is much older. So far, I have addressed cases and debates from philosophical and scientific traditions. In this last section, I address a prototypical case of altruism that comes from a religious context. This is the biblical parable of the Good Samaritan, presented in the New Testament (English Standard Version Bible, 2001, Luke 10:25-37).

The story starts with a traveler who was half-dead on a road, after being attacked by thieves. Two men, including a priest, passed by the wounded man, but ignored him, not offering any help. After them, however, a Samaritan saw the man on the road and, *feeling compassion*, "went to him and bound up his wounds, pouring on oil and wine. Then he set him on his own animal and brought him to an inn and took care of him" (English Standard Version Bible, 2001, Luke 10:34). The Good Samaritan went further and spent his own money helping the unknown man he rescued. After telling the parable, Jesus said to his disciples: "You go, and do likewise" (English Standard Version Bible, 2001, Luke 10:37).

We can find in the Bible multiple passages that address important aspects of altruism. For example, one passage says: "[l]et each of you look not only to his own interests, but also to the interests of others" (English Standard Version Bible, 2001, Philippians 2:4). We can see the concern with subtle forms of egoism that can undermine one's good deeds in another passage:

"[w]hen you give to the needy, sound no trumpet before you, as the hypocrites do in the synagogues and in the streets…. when you give to the needy, do not let your left hand know what your right hand is doing"[91] (English Standard Version Bible, 2001, Matthew 6:2-3).

The parable of the Good Samaritan can be regarded as one of the most famous prototypical cases of altruism, providing a neat illustration of its crucial features. The Samaritan did not expect any rewards from the beneficiary or others. The Samaritan did not know the man. The Samaritan did not have any obligation to help. He simply felt compassion and dedicated time and resources to help. The issue I will discuss in this section is the important role of *actions* for altruism.

Remember that, in the standard account of altruistic motivation, ultimate desires to increase the welfare of others are *necessary* and *sufficient* for altruism. Batson (2011) says that "[t]he distinction between altruism and egoism is qualitative, not quantitative. It is the ultimate goal, not the strength of the motive… that distinguishes altruistic from egoistic motivation" (p. 22). So, altruistic motivation "may evoke a variety of behaviors or no behavior at all" (Batson, 1991, p. 9). Altruistic desires can be concomitant with egoistic desires, and if egoistic desires are stronger, altruistic desires may have no behavioral output (see Sober & Wilson, 1998, p. 245).

[91] However, before this passage it is said: "[b]eware of practicing your righteousness before other people in order to be seen by them, for then you will have no reward from your Father who is in heaven" (English Standard Version Bible, 2001, Matthew 6:1). This illustrates that the concern for others can be ultimately egoistic, for one's good deeds are rewarded by God. This criticism can be applied to all instances of altruism in the Bible, since all good actions are supposed to be rewarded in the afterlife. The parable of the Good Samaritan, for example, is introduced after a lawyer asked Jesus "what shall I do to inherit eternal life?" (English Standard Version Bible, 2001, Luke 10:25). In his response, presenting the case of the Good Samaritan, Jesus does not seem concerned with the egoistic root of the inquiry. This poses a challenge for authors associating Jesus with altruism (e.g., Kaiser, 2017).

The motivational account of altruism, by focusing only on desires, rules out the need for actions. But, in doing so, I argue, it falls short of representing prototypical cases of altruism. To see the problem of neglecting actions, consider the two other men from the parable of the Good Samaritan. These men, differently from the Samaritan, did not help the man on the road. They could perfectly well have had ultimate desires to increase the welfare of the man on the road. If psychological altruism is true, and ultimate desires to help others are possible, it is possible that they had genuine ultimate desires to help. These desires, however, could have been suppressed by their egoistic desires, which happen to be stronger. According to the standard account, since altruistic motivation may not produce behaviors, the fact that the two men passing by the wounded man on the road did not act on their ultimate desire to help would not undermine the claim that their motivation was genuinely altruistic.

If the two men who denied help had ultimate desires to help, should we consider them to be altruists? Would the simple existence of such desires constitute a good criterion to decide whether we should apply the term "altruist" to someone? I believe that if someone concluded that the two men who denied help were altruists if they had ultimate desires to help, common sense would consider that this person does not know how to use the word "altruism". The action performed by the Samaritan is not a secondary, contingent aspect of his altruism, but a crucial feature of it. Merely desiring to help does not justify the use of the term "altruism".

The importance of actions for genuine cases of altruism has been discussed by other authors. Galston (1993) claims that it is reasonable to "reserve the term 'altruism' for acts in which the inner impulse to assist others comes to dominate self-regarding desires that counsel inaction" (p. 121). similarly, Schefczyk and Peacock (2010) argue that we should use "altruism" to describe cases in which other-directed desires are capable of motivating actions (p. 173).

The cases discussed in the previous sections represent prototypical cases of altruism in part due to the significant risk that agents incurred and the fact that their help was directed to people outside their close circle. But an even more basic feature is that these people not only desired to help — they *actually helped*. Oliner and Oliner (1988) discuss the differences between merely desiring to help and helping. The extensive data collected by the authors in their project shows that non-rescuers often shared with rescuers sentiments and wishes for helping, but they did not act on them (Oliner & Oliner, 1988, p. 187). What really matters, in the authors' view, is not only having a desire to increase the welfare of others, but the "feeling of responsibility for the welfare of others, including those outside their immediate familial or communal circles" (Oliner & Oliner, 1988, p. 249). The authors argue that these feelings of responsibility were more important than the mere desire to help. By neglecting the need for action, the standard account of altruistic motivation fails in reflecting ordinary altruism, as illustrated in the case of the Good Samaritan.

In this chapter, I argued that we can identify clear, prototypical cases of altruism and that these prototypical cases offer limited but legitimate access to ordinary altruism. I have also demonstrated how an account of altruism that reduces it to ultimate desires to increase the welfare of others fails in representing these prototypical cases. The prototypical cases of altruism discussed here show how altruism involves much more than ultimate desires. Expecting costs, adjusting the scope of altruism to include people from outside one's immediate circle of friends and family, and actually helping others are all key elements that make us call the cases discussed here "altruistic". By ignoring these features, the standard account of altruism fails as a candidate to accurately represent ordinary altruism.

I do not claim that what I have presented here is a fatal blow against the special status hypothesis. All I claim is that the analysis of this chapter serves as a significant piece of evidence against this hypothesis. Proponents of the special status will need to provide good evidence for why one should accept it. Until they provide such evidence, the argument presented here suffices for us to reject the special status hypothesis. Together with the criticism raised in previous chapters, this shows that the standard account of altruistic motivation is just one technical account of altruism among others. It does not deserve special treatment. So, if it proves to be an unfruitful, unhelpful account, we have good reasons to abandon it.

When I talk about altruism in informal conversations, people are often surprised to discover that the standardly accepted account of altruistic motivation does not demand actions, ignores the scope of altruism, and neglects the condition of being costly. I believe that their surprise is not an indication of their misunderstanding of altruism, but rather an indication of the distance between altruism as it is conceived in the standard account of altruistic motivation and as it is conceived in ordinary language. This chapter has offered support for this belief.

Chapter 9

Rethinking Altruism: From Desire to Virtue

9.1 The Normative Character of Altruism

In the previous chapter, I discussed many aspects of altruism that are neglected in the standard account of altruistic motivation. I concluded that the standard account fails in representing ordinary altruism. The first two sections of this chapter address other neglected features of ordinary altruism. This first section introduces the idea that ordinary altruism is a thick concept, with a normative component. *Section 9.2* defends the view that, in order to account for the normative character of ordinary altruism, we need to consider it as *context-dependent*. After this, in *Section 9.3*, I introduce an alternative account of altruism, which I call "virtue altruism". This account considers altruism to be a virtue rather than a specific mental state. *Section 9.4* offers a final criticism of the standard account, claiming that, by making altruistic motivation inaccessible, it actively discourages altruistic actions. Lastly, in *Section 9.5*, I discuss how virtue altruism avoids the challenges raised against the standard account and allows us to conceptualize altruism in a way that promotes altruistic actions.

The standard account of altruistic motivation neglects the normative character of altruism. This account is merely *descriptive*. Once we define altruistic motivation as just *any* ultimate desire to increase the welfare of others, then the label "altruistic" is deflated from its praiseworthiness. It is true that helping others is usually something praiseworthy, but in many

cases, it is not[92], and the standard account allows blameworthy desires to be altruistic. The standard account does not distinguish to whom the helping is directed, in which context, for what reason, etc. In the standard account of altruistic motivation, altruism is considered to be simply a description of certain mental states (ultimate desires) with no intrinsic moral value (see Sober & Wilson, 1998, p. 237).

However, notwithstanding its absence in the standard account of altruistic motivation, a positive moral value is a relevant feature of ordinary altruism. In the discussion of the three prototypical cases, in the previous chapter, we see that the actions described as altruistic carry an implicit approval. All three examples are morally praiseworthy, and this does not seem to be a mere coincidence. Describing something blameworthy as "altruistic" seems to be a misuse of the term. The normative character of altruism seems to be an important, fundamental feature of ordinary altruism. As Schefczyk and Peacock (2010) claim, it is surprising how the normative aspect of altruism has been neglected in the literature.

To illustrate the implicit moral approbation linked to altruism, imagine an SS officer in WW2. Imagine that, among his egoistic and evil motivations, this officer also has an ultimate desire to help his fellow Nazis. Knowing this fact about his motivational structure, would we call him an altruist or take his motivation to be genuinely altruistic? Would we call "altruistic" a slave owner who, out of an ultimate desire to help, gives a cup of water to his slave after beating him? I believe that were we to describe these people as altruists, people would question our understanding of the meaning of this word — or, worse, question our moral character.

[92] Moreover, after my analysis of the standard account in this thesis, the disconnection between this account and normativity is even clearer. In the standard account, no subjective experience nor behavior follows necessarily from altruistic motivation. Considering this, we can conclude that there are no necessary consequences to which a moral value could be intrinsically linked.

Analogously to altruism, the standard account of *egoistic* motivation is also detached from its moral value. Egoistic motivation is defined as an ultimate desire to increase one's own welfare. But when we call someone an "egoist" in everyday life, we are not merely saying that this person is doing something that aims to benefit him or herself. In colloquial language, we do not use egoism to describe just any action or desire that aims to benefit oneself. For example, we do not call the action of brushing one's teeth "egoistic", even though it aims to benefit the agent. We use the word "egoistic" when we do not consider others' interests *in the context* in which we expect one to do so. Contrasting with the standard accounts, the everyday uses of "altruism" and "egoism" carry a strong moral connotation.

At this point, we should ask: why have philosophers and scientists adopted a morally neutral definition of altruistic motivation? A plausible answer is that, by doing so, they allowed altruistic motivation to become a simpler object of study. Reducing altruism to a specific mental state allows us, in principle, to measure it without having to consider questions such as whether this particular state deserves a positive moral status. This methodological reason for detaching altruism from normativity can facilitate research, but this does not change the fact that it makes altruistic motivation something different from what it is in its ordinary use.

Batson (2011) comments on the issue of the moral status of altruism. He recognizes that it is common to attribute a positive moral value to altruism and also believes that selfishness, which is often used interchangeably with egoism, is often considered to be the very "epitome of immorality"[93] (Batson, 2011, p. 26). However, he states that attributing a moral value to altruism

[93] Batson also claims that egoistic motivation should not be equated with selfishness, for while selfishness has this strong moral connotation, egoism does not. However, we can challenge Batson's sharp distinction, here, not only mentioning the egoism is usually normative, but also remembering that the term "selfish" is used, sometimes, without any moral connotations, as in Richard Dawkins's *The Selfish Gene* (1976/2006), for example (see Sober & Wilson, 1998, p. 6).

seems to follow from an inferential mistake (Batson, 2011, p. 26; Batson, 2018, p. 20). Batson (2011) claims that some people consider self-interest as equated with the morally negative notion of selfishness, so "[i]t may seem to follow logically that if self-interest is not moral, and altruism is not self-interest, then altruism is moral" (p. 26). Batson (2011) then explains that this reasoning is flawed, for "to say that A (self-interest) is not B (moral) and that C (altruism) is not A does not mean that C is B" (p. 26).

I consider implausible Batson's hypothesis that the moral value of altruism is the consequence of a (gross) logical mistake. I do not see any good reasons to believe that this is the base of the attribution of a positive moral value to altruism. The normative character of altruism is a central feature of its ordinary use. Any account that aims to represent ordinary altruism needs to account for this normative character. In the next section, I will consider another important aspect of altruism, namely, its context-dependence.

9.2 The Context-Dependence of Altruism

The standard account of altruistic motivation reduces altruistic motivation to a discrete, context-independent mental state. This reduction makes altruistic motivation something independent of the moral/social context in which it is being used: individuals either have it or do not have it, regardless of other variables. This section contends that "altruism", in its ordinary use, is context-dependent. That is, the relevant features determining what counts as altruistic vary according to the context. To explain this idea, I will start by discussing a feature of altruism that has not received much attention in the thesis to this point, namely, the notion of the "welfare of others".

Remember that, in the standard account, the welfare of others is that which altruistic motivation aims to increase. As I discussed in Chapter 2, Kitcher (2011) explains that there is a distinction between paternalistic and non-paternalistic altruism. In paternalistic altruism, agents aim to increase what the agents themselves believe to increase others' welfare. In non-paternalistic altruism, agents aim to increase what the beneficiaries themselves believe will increase their own welfare. These two accounts of the welfare of others will often diverge[94]. So, when they diverge, which of the two approaches should we adopt?

Although the standard account of altruistic motivation is not explicit regarding which approach one should adopt, the authors discussing psychological altruism usually adopt paternalistic altruism. Sober and Wilson (1998) claim that "altruists have ultimate desires concerning what they think will be good for others" (p. 230). Batson (2018) reflects the same opinion when he says that "to perceive another in need involves seeing a discrepancy between the other's current state and what you think is good for him or her" (p. 189). Regardless of whether the standard account remains neutral in this issue or whether it is always paternalistic, what is important is that, in both cases, it is not sensitive to the context. Following Kitcher (2010, 2011), I will defend that the decision between paternalistic and non-paternalistic should be made taking into account the context.

Kitcher (2010) claims that, "in some circumstances, only alignment with the wishes [of others] counts as genuine altruism" (p. 132). But, he continues, in other circumstances, only alignment with what we believe to be the best for them is altruistic. So, the response proposed by

[94] A parallel distinction is also possible in egoism: there is a difference between personal preference and personal benefit (see Kavka, 1986, p. 40). What counts as a self-benefit vary depending on who is judging. Thus, there is a distinction between a paternalistic and a non-paternalistic view of self-benefit.

Kitcher, which I endorse, is that neither approach is appropriate for all cases — they depend on the context.

To see the context-dependence mentioned above, consider an example. Imagine a thirsty child who happens to find a cleaning product that smells just like apple juice. If this child wants to drink this product, based on his or her false belief that this would increase her welfare, the appropriate action is to go against his or her preference and adopt *paternalistic* altruism. In other cases, however, the appropriate thing to do is to respect others' preferences. For example, I might struggle to identify the benefits of living in an isolated tribe in the middle of the Pacific, without access to modern medicine. But if the people living there prefer this life, I should accept and respect their take on what is best for them. Imposing on them what I believe to be the best for them would hardly be considered altruistic. As Kitcher (2010) points out, in some cases, we should simply respect others' preferences (p. 132). The decision between paternalistic and non-paternalistic altruism cannot be established independently of the context in which we are making the decision.

Remember that the standard account of altruistic motivation is very clear regarding excluding instrumental desires from the scope of altruism. This seems appropriate, as it rules out some cases that would be odd to call "altruistic". However, it seems equally odd to call altruistic someone who gives a child a cup of disinfectant because the child thinks this would greatly increase her welfare. The same is true for someone with paternalistic inclinations who imposes a western way of life over Aboriginal people under the argument that "this is for their own good". The standard account of altruistic motivation ignores these differences, taking all of them to be altruists if they have ultimate desires to help. This fails, however, in recognizing the importance of the context in our attributions of altruism.

It is widely accepted in the debate on psychological altruism that knowing the helping behavior is not sufficient to tell us whether one has altruistic motivation (see Peacock et al., 2005). What I am proposing here is that, analogously to that, knowing whether one has a *specific state of mind* (namely, an ultimate desire to increase the welfare of others) is also not sufficient to determine whether we should consider one's motivation altruistic. The criteria used in ordinary language to decide whether to use the label "altruistic" or not is not merely a matter of knowing the kind of desire one has.

The very same action, following from the very same mental states, could be considered altruistic in one context and not in another. To illustrate this, imagine the following situation. You are in your house late at night and your neighbor asks if it is possible to borrow your car. His car is broken, and he needs to take someone to the hospital. Imagine that you reasoned that you want to help, but you also do not want to be financially impacted by the problem of someone unrelated to you. So, you offer your neighbor your car, adding the condition that he should return the car with the same amount of gas in the tank and that he would be responsible for any damage caused to the car. Are you an altruist? It depends on the context.

If you were in a social context where no one ever lends their cars, you could be taken as an altruist for helping. However, if you happen to be in a place where everyone is very solicitous, lending their cars to whoever asks, the fact that you asked for a full tank could qualify you as quite an egoistic person, someone concerned with a few dollars over another person's welfare in a moment of emergency. The crucial point to notice in the comparison between these two contexts is that the same person, engaging in the same behavior, and *with the same mental states*, would be considered an altruist in one context and not in the other. This illustrates how an

account of altruistic motivation, defined as a specific context-independent mental state, is unsatisfactory.

In this section, I focused my discussion on the welfare of others. But there are many variables at play when we consider whether to use the term "altruism". The previous chapter discussed a series of features that, depending on the context, will be more or less important. As Schefczyk and Peacock (2010) discuss, the contexts in which we will determine how and when to use altruism are complex and involves many factors.

> Whether one is in the habit of ascribing ulterior motives to an apparent altruist depends partly on social conventions regarding the appropriate exercise and expression of cynicism or benevolence regarding the motives of others. Some people or some cultures might be more cynical about the claims of altruism than others, and the determinants of cynicism, and not just the allegedly ulterior motives of the altruist, are worthy of investigation. (Schefczyk & Peacock, 2010, p. 176)

Bentham (1789/2000) discusses a similar issue in his discussion of benevolence:

> A man who has set a town on fire is apprehended and committed: out of regard or compassion for him, you help him to break prison. In this case the generality of people will probably scarcely know whether to condemn your motive or to applaud it: those who condemn your conduct, will be disposed rather to impute it to some other motive: if they style it benevolence or compassion, they will be for prefixing an epithet, and calling it false benevolence or false compassion. (p. 91)

Those who disapprove of the action will resist using the term "benevolence", as this term denotes approbation. The same goes, I claim, for altruism. Ordinary altruism does not depend only on the mental states and actions of agents, but also on the moral worth of these states and actions. In the next section, I propose a different approach to altruism that aims to account for its normative character and its context-dependence.

9.3 Virtue Altruism: An Alternative to Rethink Altruism

In Chapter 8, when discussing the work of Oliner and Oliner (1988), I briefly mentioned the idea of an *altruistic personality*. As Penner et al. (2014) claim, the idea of measuring stable altruistic traits in individuals has its origins in psychology in the 1980s. Before that, psychologists considered propensities to prosocial behaviors as mainly *situationally* determined. But as researchers identified several personality traits that reliably predicted prosocial action, the idea that some people have altruistic personalities became more popular.

The main traits considered to constitute the altruistic personality are helpfulness, defined as other-directed empathy and "the tendency to provide help to needy individuals and groups of individuals" (Penner et al., 2014, p. 149). An individual with an altruistic personality is described as someone with "higher standards of justice, social responsibility, modes of moral reasoning, who is more empathic to the feelings of others" (Feigin et al., 2014, p. 4). The altruistic personality is considered to be "a relatively enduring predisposition to act selflessly on behalf of others, which develops early in life" (Oliner & Oliner, 1988, p. 3). The idea of an altruistic personality opens a different perspective on altruism: rather than conceptualizing altruism as a discrete mental state, we may conceptualize it as a personality trait. The altruistic personality involves a complex set of elements, both cognitive and conative, and is not reduced to a single motivational state.

In this section, I will propose an account of altruism that builds on the idea of an altruistic personality but that also integrates the context-sensitivity of altruism and its moral dimension. The proposal is to conceive altruism as a *virtue*. I will call this account "*virtue altruism*" and defend the claim that it avoids some of the problems that I have been raising against the standard

account of altruistic motivation. I start by discussing virtues and explaining how we can make sense of altruism as a virtue. After this, I will briefly present some of the benefits that follow if we adopt virtue altruism: it offers a way of accounting for the features of ordinary altruism discussed in the previous chapter; it accommodates the descriptive and normative characters of altruism; and it integrates the context-sensitive character of altruism, preserving its moral value.

Aristotle presents virtues as a mean between extremes (Rachels & Rachels, 2015, p. 162). To illustrate this property of virtues, consider the Aristotelian virtue of courage. "[T]he person who avoids and fears everything, never standing his ground, becomes cowardly, while he who fears nothing, but confronts every danger, becomes rash" (Aristotle, 2000, p. 25). Courage is a mean between a deficiency and an excess: cowardice and rashness, respectively. A courageous person is not afraid of doing the right thing when facing danger, but she is also not unnecessarily risking her life and wellbeing. Virtues "are ruined by excess and deficiency" (Aristotle, 2000, p. 25). A courageous person navigates a sea of cowardice and immoderate risk-taking, aiming to find the right balance.

A first question, if we plan to characterize altruism as a virtue, is whether we can make sense of virtue altruism as the mean between two extremes. The usual interpretation is that altruism is the opposite of egoism, but it is not clear what "opposite" means here. One interpretation could be that altruism and egoism are two extremes in the same way that cowardice and rashness are extremes. If this is so, then both altruism and egoism are vicious traits to be avoided. But I propose a different way of thinking about the opposition between altruism and egoism. Altruism is better conceived not as an extreme (as an excess or a deficiency) but as the mean between two vicious extremes. While courage is the mean between cowardice and rashness, altruism is the mean between egoism and *self-abnegation*.

The idea of altruism as something between egoism and self-abnegation has been articulated by some authors. Kitcher (2010) proposes a scale that goes from egoism to self-abnegation (p. 127). This view of altruism as something that comes in degrees is consistent with the idea of virtues (see Hursthouse & Pettigrove, 2018). Kitcher (2010) says, for example, that, in cases in which the deviation from the egoistic extreme is too low, one is altruist only "in a very modest sense" (p. 127). Spencer also discusses a similar idea in his *Data of Ethics* (1862/1883). He claims that altruistic behaviors, if excessive, should not be considered altruistic at all — when pushed to an extreme, unselfishness becomes selfish (Spencer, 1862/1883, p. 196). Galston (1993) corroborates this view, claiming that "certain kinds of concern for others are rooted in a lack of concern for, or undervaluing of, oneself that is more nearly a *vice than a virtue [emphasis added]*" (p. 120).

Self-abnegation, specifically in the sense of ignoring one's own welfare completely when aiming to benefit others, is not praiseworthy. Altruism, if it is morally praiseworthy, is not an absolute, immoderate self-destructing behavior but an adequate concern for others. Self-abnegation is an absolute concern for others with no regard for oneself, but in virtue altruism, one does not completely disregard oneself.

If altruism and egoism are two extremes, then both are vicious. However, this would be inconsistent with the positive moral value of altruism. By contrast, if we depict altruism as the mean between excesses, we explain its opposition to egoism and account for its moral value. The view of altruism as the mean between excesses is consistent with the discussion in this and the previous chapters. But if we accept that altruism should not be equated with self-abnegation, one can raise an objection. Altruism does not seem to be simply at the mid-point between egoism and self-abnegation: it seems more distant from egoism than it is from self-abnegation. However, this

is a problem that we can address by better understanding the view of a mean between extremes in Aristotelian virtue theory.

"In some cases," says Aristotle (2000), "the deficiency is more opposed to the mean than is the excess, in others the excess is more opposed than the deficiency; for example, it is not rashness, the excess, which is more opposed to courage, but cowardice" (p. 35). Aristotle explains this variation as a consequence of our natural *inclinations* to one extreme. "It is the things to which we ourselves are naturally more inclined that appear more contrary to the mean; for example, we are naturally more inclined to pleasures, and are therefore more prone to intemperance than self-discipline" (Aristotle, 2000, p. 35). This is why courage seems "opposite" to cowardice. Since we are more prone to cowardice than to rashness, courage pushes us towards the rashness extreme.

Considering the idea that virtues compensate for our natural inclinations, we can explain why altruism seems more distant from egoism than it is from self-abnegation. Our nature pushes us to the egoistic side. So, when we see a case of altruism it seems something remarkably distant from egoism. "[T]he mean state is in every case to be praised, but that sometimes we must incline towards the excess, sometimes towards the deficiency, because in this way we shall most easily hit the mean, namely, what is good" (Aristotle, 2000, p. 36). The virtue of altruism concerns the care for others, but not an excessive one. Virtue altruism avoids both selfishly motivated acts of helping and acts that disregard oneself.

Before discussing the benefits of adopting virtue altruism, I should state a significant limitation of this account. Remember that one of the problems of the standard account of altruistic motivation is the difficulty in identifying cases of altruistic motivation. Despite its advantages, virtue altruism faces a similar difficulty. The language of virtues is inherently vague

and hard to grasp. So, although virtue altruism reflects the ordinary use of altruism, its multifaceted, complex, and normative character makes its use in scientific research difficult. I will make two comments about this limitation of virtue altruism.

The first comment is that one can accept my criticism of the standard account of altruistic motivation without having to adopt virtue altruism. This criticism and my proposal of conceiving altruism as a virtue should be distinguished. I accept that, for certain purposes, virtue altruism will not be the best concept. The technical accounts of altruism have an important place and I did not argue that we should substitute them all to virtue altruism. Evolutionary altruism, for example, is a well-established concept and there is no reason to change it based on what I argued here. What we should be aware of, however, is how these technical accounts fail to reflect the normative character of ordinary altruism. So, if we want to use the term "altruism" in a way that reflects ordinary altruism, with its morally praiseworthy character, we should prefer virtue altruism over the other options. In particular, virtue altruism offers a better account of altruism for moral philosophy. The moral debate about altruism would be better articulated if instead of focusing on ultimate desires we focused on virtues.

The second comment I want to make about the limitation of virtue altruism is that, differently from the standard account, the difficulties in identifying a case of virtue altruism follow from the complex nature of altruistic motivation, not from the decision of using a specific problematic concept. In the case of the standard account, the challenge of identifying altruistic motivation follows from the problematic reduction of altruistic motivation to "ultimate desires". In the case of virtue altruism, the difficulty follows from the inherently complex nature of the phenomenon being described. Ordinary altruism refers to a complex phenomenon, which needs to take into account many variables, and this complexity is reflected in virtue altruism. So,

although both accounts offer similar challenges when it comes to identifying cases of altruistic motivation, I believe that virtue altruism offers good justification for its difficulty.

Consider now some advantages of adopting virtue altruism. Virtues are neither defined solely in terms of behaviors nor in terms of motivations. Virtues are stable traits of character that involve not only a specific mental state, but a complex set of dispositions to act and feel in certain ways. An honest person, for example, is not merely behaving honestly or merely wishing to behave honestly, but someone who is honest. Being honest involves behaving honestly, desiring to be honest, feeling good when acting honestly, and so on (see Hursthouse & Pettigrove, 2018). Importantly, virtues are also morally praiseworthy. As Rachels and Rachels (2015) comment, "virtue is a commendable trait of character manifested in habitual action" (p. 161). So, virtue altruism has a descriptive dimension, including a set of personality traits, and a normative dimension, which value this as morally positive. Thus, virtue altruism offers a way of accommodating both the descriptive and the normative characters of altruism, which were shown to be neglected by the standard account of altruistic motivation.

Differently from the standard account, which reduces altruistic motivation to a specific mental state, virtue altruism is not reducible to a discrete mental state. Virtue altruism is instantiated in actions that benefit others, motivated by desires to help *relevant* others, promoting the welfare of these relevant others to the right degree, in the right way, for the right reasons, and in the right context. All of these variables are constitutive of virtue altruism. By integrating these aspects of altruism, virtue altruism could reflect the features of ordinary altruism illustrated in the prototypical cases addressed in the previous chapter. Moreover, virtue altruism also helps to close the gap between motivational and behavioral altruism by characterizing altruism as a virtue with both motivational and behavioral components.

If I am correct in claiming that virtue altruism offers a good representation of ordinary altruism, then we can understand why the scientific and philosophical approaches are misguided when it comes to representing ordinary altruism. To illustrate this, imagine a hypothetical scenario where *courage*, rather than altruism, is considered to be a motivational state. Imagine that psychologists decided to make an empirical quest to identifying cases of courageous motivation. They propose experiments to answer "the courage question": do humans have genuine courageous motivation? Assuming that courage should be observable, they establish that courage should be reduced to a specific mental state, and propose experiments to identify such a state. This, of course, would misrepresent courage, since it would neglect the multiple variables involved in this virtue. It is true that courage depends on certain psychological states, such as confidence, self-control, etc. But we cannot *reduce* courage to one of these states. The same is true for virtue altruism. Reducing altruistic motivation to a specific mental state (ultimate desires) is the central problem underlying the debate about psychological altruism, and this problem is avoided if we adopt virtue altruism.

Virtue altruism also offers a way to account for the context-dependence of altruism. As is the case for other virtues, the degree of importance of each of these variables is sensitive to the context and will be determined differently in each case. As Aristotle (2000) says when discussing the nature of virtues, "agents must always look at what is appropriate in each case as it happens, as do doctors and navigators" (p. 25). The nature of virtues is such that we cannot rely on fixed rules in order to know what the virtuous thing to do is. Issues such as the problem of choosing between paternalistic and non-paternalistic altruism are not determined *a priori*, but decided in each situation.

In this section, I introduced the notion of virtue altruism and presented some benefits that it offers. Virtue altruism articulates a similar idea to that of the altruistic personality. In Oliner and Oliner's (1988) discussion on the profile of the rescuers, one of the conclusions they reach is that a major cause for these people's extraordinary acts of care was their internalized values. These values — usually traced back to their parental education — formed a character, a profile with stable dispositions to care for others. Hence, the authors say, "their actions may appear impulsive, without due consideration of consequences. In fact, however, they are merely the extension of a characteristic style of relating developed over the years" (Oliner & Oliner, 1988, p. 251). If debates over altruism are focused solely on ultimate desires, we miss other components of motivation, such as the stable personality traits, considered by Oliner and Oliner (1988) to be fundamental to altruism. Virtue altruism depicts altruism as an appropriate, praiseworthy concern for others, allowing us to conceptualize altruism without needing to invoke ultimate desires.

9.4 The Standard Account Discourages Altruistic Actions

This thesis has raised a series of arguments against the standard account of altruistic motivation. In this section, I raise a distinct kind of argument. I will argue that conceiving altruistic motivation as an ultimate desire to increase the welfare of others actually discourages helping actions in the real world. Assuming that having more altruistic behaviors in the world is a morally good thing, the discouragement of altruistic actions caused by the standard account of altruistic motivation is a morally negative consequence. Here I will articulate a *moral* argument against the standard account of altruistic motivation. This argument poses yet another reason for

rethinking altruistic motivation. In the next section, I explain how virtue altruism can avoid this problem and actively encourage altruistic actions.

The argument presented in this section is divided into four parts. Here is a quick summary. In the first part, I claim that believing that our motivation to help is altruistic typically produces an extra motivation for us to perform the helping behavior. This is because altruistic motivation is typically considered morally praiseworthy, and people tend to prefer to act in accordance with their moral principles. In the second and third parts, I claim that, since ultimate desires are not accessible, people who characterize altruism in terms of ultimate desires are unlikely to believe that their motivation is altruistic. Thus, they will not obtain the extra motivation to help that is produced by believing that one's motivation is altruistic. Thus, they will be less likely to perform the helping behavior. Finally, in the fourth part, I claim that such a reduction in the likelihood of producing helping behaviors is a morally negative outcome.

The *first* part of the argument says that,

> For any agent A considering performing a helping behavior B,
>
> (P1) if A believes that her helping behavior B is motivated by altruistic motivation, then there is an increase in the likelihood that she will also believe that her helping behavior B will be morally praiseworthy;
>
> (P2) if A believes that her helping behavior B will be morally praiseworthy, then A will have an *extra motivation M* to perform helping behavior B.
>
> (C1) if A believes that her helping behavior B is motivated by altruistic motivation, then there is an increase in the likelihood that A will have an *extra motivation* M to perform helping behavior B. (*From P1, P2, by hypothetical syllogism*).

First of all, notice that (P1) is *not* assuming that helping behavior is morally praiseworthy *only when* motivated by altruistic motivation. Although some authors seem to endorse such a

view (e.g., Shafer-Landau, 2012), my argument does not rely on such a strong assumption. I assume simply that people *usually* believe that actions are *more* morally praiseworthy when they are motivated by altruistic motivation. Donating one hundred dollars in order to avoid taxes may be praiseworthy in itself, but it is less praiseworthy than donating one hundred dollars out of a desire to improve the quality of life of others. All that (P1) assumes is that the view of altruistic motivation as morally praiseworthy is a commonly held view. Since people *often* value altruistic motivation as morally praiseworthy, then it is more likely that agent A will believe that her helping behavior B is morally praiseworthy.

Consider now the second premise. Why do people have such an extra motivation to perform a behavior they believe to be morally praiseworthy? Simple: because most people *desire* to act morally. They may different reasons for doing so, such as aiming to get the social or subjective rewards it elicits, or in order to be consistent with the principles that they hold, etc. What matters here is that, in general, people have moral standards and they have a preference for acting accordingly to these moral standards. When they believe that the motivation for a given action is morally praiseworthy, they have further reasons to perform this action[95]. In other words, in addition to the motivation one has to perform a given behavior, if one believes that this behavior is morally praiseworthy, one has an extra motivation to perform it.

For most people, the moral praiseworthiness of an action constitutes an extra motive for them to perform such an action. But how strong will this preference be, and what other things

[95] Reasons can play different roles. We can give reasons to explain, to justify, or to motivate (see Alvarez, 2009). I should be clear that the reasons I am mentioning here are neither explanatory nor justificatory. I am not claiming that acting moved by altruistic motivation is right and therefore one is justified to perform it. The claim is that, due to the beliefs people happen to have in general, believing in the praiseworthiness of an action makes them more likely to perform it. It is a *motivational* reason. These people have more reasons in their decision-making to perform the helping behavior.

will be valued even more, thus precluding one from acting morally? That is a difficult question. The preference to act morally might be very weak for some individuals. For these individuals, egoistic desires will easily undermine this moral preference. But the crucial point for us here is that, even for these very selfish individuals, who can easily depart from their moral principles, if they can choose two equally advantageous and equally costly options, they are more likely to choose the one that is in accordance with their moral principles. This is the assumption we need to make in order to support (P2).

It should be noted that the view of altruistic motivation as something morally praiseworthy finds support in the main normative theories. If one's ethical judgments are based on virtue ethics, for example, the existence of an appropriate motivation can easily be seen as something that makes an action more praiseworthy. We can say, following Aristotle (2000), that moral actions should be done in the right way, at the right time, and with the right *motivation*. So, if we think of altruism as a virtue, virtue ethics can account for the moral value of altruistic motivation. Likewise, a utilitarian perspective, although interested only in the consequences of actions, can still value altruistic motivation. It can do so by considering this motivation as a device that tends to produce good actions in the world. Finally, an approach based on Kantian ethics might also account for the moral value of altruistic motivation. A Kantian approach could indeed claim that actions should be valued on the basis of our reasons for doing them, not on our desires: truly praiseworthy moral actions are not conditional on our affective states, but to reason. However, some authors have argued that the normative status of altruistic motivation may be compatible with Kant's moral theory, for it can be conceived as an "imperfect duty" (see Schefczyk & Peacock, 2010, p. 182).

Consider now the *second* part of the argument, which considers what follows if one adopts the standard account of altruistic motivation.

> (P3) If A defines altruistic motivation as an ultimate desire to increase the welfare of others, then there is a reduction in the likelihood that A will believe that she has altruistic motivation.
>
> (P4) If there is a reduction in the likelihood that A will believe that she has altruistic motivation, then there is a reduction in the likelihood that A will have the extra motivation M.
>
> (C2) If A defines altruistic motivation as an ultimate desire to increase the welfare of others, then there is a reduction in the likelihood that A will have the extra motivation M. (*From P3, P4, by hypothetical syllogism*).

To see why (P3) follows, we just need to remember that ultimate desires are virtually inaccessible. For any desire we have, it is possible that it is merely instrumental to an egoistic desire. Whether a desire is ultimate is something that can escape both introspection and the observation of behaviors. Therefore, if we define altruistic motivation as an ultimate desire to increase the welfare of others, then it is virtually impossible for us to know whether we have altruistic motivation. Thus, it becomes very hard for us to believe that we have altruistic motivation[96].

To understand why we should accept (P4), we need to remember the first part of the argument. In (C1), it is said that "if A believes that her helping behavior B is motivated by altruistic motivation, then A will have an *extra motivation* M to perform helping behavior B". So, considering this, the more we reduce the likelihood of A believing that her motivation is

[96] One can certainly form an unjustified belief that one's ultimate desire to help is ultimate. This belief might even turn out to be true. But the better one understands the challenges involved in accessing altruistic motivation, the less likely it becomes for one to form such belief. Remember that the argument is only saying that there is a reduction in the likelihood that A will believe that she has altruistic motivation, not that this outcome is impossible.

altruistic, the more we reduce the likelihood that she will obtain the extra motivation M. The conclusion shows that the simple definition of altruistic motivation as an ultimate desire reduces the likelihood that A will have the extra motivation M. The *third* part of the argument makes it explicit that, by precluding one from obtaining extra motivation M, we also reduce one's likelihood of performing helping behaviors.

> (P5) If there is a reduction in the likelihood that A will have the extra motivation M, then there is a reduction in the likelihood that A will perform helping behavior B.
>
> (C3) If A defines altruistic motivation as an ultimate desire to increase the welfare of others, then there is a reduction in the likelihood that A will perform helping behavior B. (*From C2, P5, by hypothetical syllogism*).

The reason for accepting (P5) is simple. The extra motivation M is a causal factor making the helping behavior B more likely to occur. All things being equal, an individual with the extra motivation M is more likely to perform helping behavior B than an individual without the extra motivation M. The less likely it is for A to have extra motivation M, the less likely it is for A to perform helping behavior B. A final step needs to be stated in order to make the moral claim of the argument. This is presented in the *fourth* part of the argument.

> (P6) A reduction in the likelihood of A performing helping behavior B is a morally negative effect.
>
> (C4) Therefore, if A defines altruistic motivation as an ultimate desire to increase the welfare of others, then a morally negative effect is produced. (*From C3, P6, by hypothetical syllogism*).

Perhaps the most controversial premise in my argument is (P6). This premise is based on an assumption that I believe to be reasonable: it is good to have more helping behaviors of the kind that is often considered to be motivated by altruistic motivation (behaviors that are at least

not moved by immediate rewards, for example). Having more people helping each other in the world is a good thing. Of course, helping behaviors can produce negative outcomes, but the claim is that, *in general*, they tend to produce more good than harm. I believe that the moral assumption underlying (P6) is a fairly common view, which is also compatible with the main normative theories.

Of course, some readers might not share the assumption above, in which case my argument will no longer have the moral implication I claim. Nevertheless, these readers can still accept that the argument shows a practical effect of the standard account of altruistic motivation, namely, that it discourages helping behavior. Assuming the importance of helping behaviors in social structures, the reduction of helping will have significant social implications that also deserve to be considered. In the following section, I discuss how virtue altruism overcomes the problem raised in this section and other similar issues that follow from the standard account of altruistic motivation.

9.5 Pursuing an Altruistic Life

One of the goals of Oliner and Oliner (1988) in their investigation of the rescuers in WW2 was to understand what attributes these helpers have, so that one could "deliberately cultivate them" (p. xviii). This is a goal that, I believe, is shared by most people: we want to cultivate our altruistic dispositions. In the previous section, I proposed an argument to show that the standard account of altruistic motivation discourages us from doing so. In this section, I extend this discussion, showing other ways in which the standard account precludes altruistic actions. But in this section, I will not only point out the problems, but show how virtue altruism

offers ways of avoiding these problems, thus promoting the creation of more altruistic actions in the world.

The argument presented in the previous section identifies a problem: since the standard account defines altruism in a way that makes it inaccessible for agents, one can never know when one's motivation is altruistic. I explained how this reduces the chances of individuals producing altruistic actions. Now, consider how virtue altruism offers a solution to this problem.

In the standard account, the distinction between altruistic and egoistic motivation is qualitative, not quantitative (Batson, 2011, p. 22). Altruistic motivation is either completely present or completely absent. However, if we adopt virtue altruism, then altruistic motivation has different *degrees*. We are more altruistic or less altruistic. So, an imperfect degree of altruism would already be altruism to some degree. This allows us to *believe* we have altruistic motivation, even if such a motivation is imperfect. If we do so, we can have the extra motivation M, discussed in the previous section. Virtue altruism avoids the discouragement of altruistic actions that follows if we adopt the standard account of altruism.

Virtue altruism avoids the problem raised in the previous section. This gives us reasons to adopt virtue altruism. But now I will consider another reason to do so, which I believe to be even stronger: virtue altruism offers a way of thinking about altruism that allows one to actively promote one's own altruistic dispositions, while the standard account does not. I will first explain why the standard account precludes one from promoting one's own altruistic actions and then explain how virtue altruism avoids this problem.

A fundamental problem with the standard account is that it does not make clear how and whether ultimate desires can be deliberately created. Imagine, for example, that I realize that having altruistic motivation is something good for myself, which I should pursue if I want to

have a happy and fulfilling life. There are at least *two problems*. The first problem is that it is not clear how one can create ultimate desires to help others. In the discussion about psychological altruism, these desires seem to be out of reach. For example, authors talk about these desires as evolved traits. If they are evolved traits, then there seems to be little one can do to have more of them[97]. There is no clear indication of what sort of thing one could do in order to create genuine altruistic motivation.

A second problem for someone aiming to cultivate altruistic motivation is that, if one accepts the standard account, one might face a fundamental contradiction. This is so because if one desires to obtain ultimate desires to help others, one will instantly fail, for these desires would be *instrumental* to one's desire to obtain these altruistic desires. If altruistic desires are ultimate desires to increase the welfare of others, then one's efforts to create altruistic desires seem to be *self-defeating*, for they would be derived from "egoistic" desires. The more one understands that pursuing a life in which we care for others makes life worth living, the more one has personal reasons to care for others. But the more one has personal reasons to care for others, the more likely that desires to help will be instrumental.

The problems of the standard account follow because it conceives altruistic motivation as something opposed to *any* self-benefit. The definition of "egoistic motivation" as an ultimate desire to increase one's own welfare is too broad. When we want to rule out egoistic motivation from the set of altruistic actions, we are not concerned about ruling out *all* sorts of actions that aim to benefit oneself. This problem is avoided if we think in terms of virtues. In virtue altruism, the role of personal benefits, and even pleasure, can be understood in a different way.

[97] Remember that, in my discussion, I am adopting what in Chapter 3 I called the "hard-access view", in which ultimate desires are few, hard to access, and hard to be produced. This is the standard approach in the literature.

Aristotle (2000) claims that virtues are pleasant, and "the person who does not enjoy noble actions is not good. For no one would call a person just if he did not enjoy acting justly, or generous if he did not enjoy generous actions" (p. 14). The joy that we feel when we see that our help actually made someone better off is not a sign of egoism. An altruistic education passes through learning how to *enjoy* being altruistic.

Recognizing the enjoyment of altruism is also consistent with the views of Comte (1851/1875, Vol. 1), who considered benevolent emotions to be the "sweetest to experience" (p. 74). This view is also shared by some of the modern authors who opposed egoistic accounts of human nature (Butler, 1726/2006, p. 117; Shaftesbury, 1711/2001, Vol. 2, p. 58; see also Grote, 2010). Virtue altruism offers a way of integrating the pleasure of pursuing the good of others without undermining the claim that this is a genuine case of altruism. If we adopt the standard account, it seems that genuine altruistic motivation is less likely to occur in individuals who recognize the benefit of helping others and enjoy doing so. By contrast, if we adopt virtue altruism, enjoying helping others is part of what it is to be a genuine altruist.

In this section, I discussed some ways in which virtue altruism offers an alternative to the standard account of altruistic motivation, explaining the advantages of conceptualizing altruism as a virtue[98]. I truly believe that having more altruistic actions in the world is a good thing. But more than believing in the value of having more altruistic actions, I also believe that there is

[98] Virtue altruism is not the only alternative to the standard account. I explained, in Chapter 2, the notions of preference altruism and Kitcher's altruistic motivation. Other authors have also proposed attempts to articulate altruistic motivation beyond ultimate desires. Piccinni and Schulz (2018, 2019) suggest that we could focus on how desires are *produced* rather than on their content. Clavien (2012) proposes moving away from ultimate desires towards an *emotional* account of altruistic motivation. Unfortunately, I did not have space to discuss all these views. However, the criticism that I raised against the standard account gives us more reasons to seriously consider these alternative accounts of altruism.

value in developing an altruistic character. If we conceptualize altruism as a virtue, we are not mainly concerned with the question of whether a particular motivation is truly altruistic or not. Instead, we are more concerned with developing an altruistic personality, integrating dispositions to help others into our character. This involves the capacity to care for others, but also the capacity to enjoy doing so, the capacity to do so in a way that does not undermine other virtues, and so on. Having dispositions to care for others beyond our close circle, accepting some personal costs, and integrating these dispositions as part of one's personality is an effective way of producing a life worth living[99].

[99] For a review on empirical evidence that may support this claim, see Post (2005).

Chapter 10

Conclusion

In his *Does Altruism Exist?* (2015), David Sloan Wilson reveals his concerns regarding the standard account of altruistic motivation. When commenting on the process of writing *Unto Others* (1998), he says that, as a biologist, he was mainly in charge of the first part of the book, dedicated to evolutionary altruism, while Elliott Sober was in charge of the part focused on psychological altruism. Wilson says that, as he learned more about the philosophers' and psychologists' approaches to altruism, he could not avoid thinking that something was going wrong. He thought that "[s]ticking to distinctions such as hedonism, egoism, and altruism seemed antiquated, since these concepts predated not only evolutionary theory but also the emergence of psychology as a science" (Wilson, 2015, pp. 60-61). Wilson went further:

> Much of the recent philosophical and psychological literature seemed like a parlor game in which proponents of selfishness described a hypothetical psychological mechanism that did not count as altruistic according to their criteria, but which produced apparently altruistic behavior…. Proponents of altruism were then required to disprove the hypothetical selfish mechanism…. After a few rounds of playing this game, the selfish psychological mechanisms being discussed were *virtually identical to altruistic psychological mechanisms in their behavioral manifestations* [emphasis added], which was why cleverness was required on the part of philosophers to tease them apart…. The more cleverness was required, the less I cared about the outcome any more than I care whether someone pays me by cash or check. (Wilson, 2015, p. 61)

It was not my goal in this thesis to argue that ultimate desires to increase the welfare of others do not exist. These desires might well exist. What I do argue, however, is that their existence is not supported by the main arguments present in the literature. More importantly, I argued that we do not have good reasons to take these ultimate altruistic desires as the sole

criterion when defining altruistic motivation. I have provided multiple reasons for us to think, like Wilson (2015), that something has gone wrong in the contemporary way of thinking about altruistic motivation.

At this point, we can question why the view of altruistic motivation as an ultimate altruistic desire has been taken as the standard account. I believe that one of the reasons is that the pursuit of an ultimate desire as the true cause of behavior reproduces the approach of searching for ultimate causes, which is a familiar idea in philosophy. Sober and Wilson (1998) claim that, if the chains of desires "cannot circle back on themselves, and if people don't have infinitely many desires, then these chains must be finite and each must trace back to a first member" (p. 350). But while this concern for finding the ultimate causes of behaviors is legitimate, we should question the legitimacy of narrowing down this search specifically to ultimate desires, while ignoring other factors. Surely, an ultimate desire is ultimate in the *chain of desires*. But an ultimate desire is not necessarily the relevant ultimate cause in the chain of *motivational states*. In the same way that an instrumental desire is conditional to other desires, ultimate desires can be conditional to other mental states, such as emotions (see Clavien, 2012). Why not consider other variables, such as the origins of ultimate desires (see Piccinini & Schulz, 2019), as equally important criteria for what counts as altruistic?

Schefczyk and Peacock (2010) claim that "[t]he time-honoured problem whether a person enjoys altruistic acts because she has other-regarding goals or whether she has other-regarding goals because she enjoys altruistic acts, is irrelevant" (p. 176). I do not go as far as to affirm that this problem is always irrelevant. Whether our desires to help others are ultimate is *one* of the variables in altruistic motivation. However, in my view, ultimate desires should not be considered the *main* feature of altruistic motivation, let alone its *only* feature. I do not see reasons

why the debate about altruistic motivation should be reduced to the quasi-metaphysical question about whether one's motivational causal chain has its basis in an ultimate (unconscious) desire to improve one's own welfare[100]. Other variables, such as cost, scope, and motivational strength, for example, should also play a role in determining what counts as altruistic motivation. Reducing altruistic motivation to ultimate desires is an arbitrary, unjustified, and potentially harmful choice done by some researchers in recent history.

In this thesis, I argued that the standard account of altruistic motivation (1) is not a useful or fruitful notion for scientific research; (2) diverges from the historical uses of the term, including the original use proposed by Comte; (3) does not represent ordinary altruism; and (4) neglects the context-dependence of altruism and does not account for its normative dimension. Notice, however, that the fact that a given account of altruism is also subjected to one of the four criticisms above does not undermine its legitimacy. For example, we can apply (2-4) to evolutionary altruism. However, evolutionary altruism is a very useful account of altruism, so we cannot apply (1) to this account. Does the fact that (2-4) applies to evolutionary altruism have any implication for the legitimacy of this account? No, it does not. Evolutionary altruism is a technical account and, as long as it avoids (1), we should keep it as part of the scientific vocabulary. Researchers in different areas, with different goals, can use different accounts of altruistic motivation, and that is good for research. However, considering (1-4), I believe that there is no research context in which the standard account of altruistic motivation appears as the best alternative.

[100] Notice that this idea is also coherent with the historical accounts discussed in the previous chapter. There, we saw authors concerned with excluding deliberate egoism. From Shaftesbury to Comte, authors have considered this to be the relevant form of egoism threatening our accounts of selflessness, not the obscure notion of ultimate egoistic desires.

This thesis also offers some important implications for moral philosophy. The view that altruistic motivation is a *requirement* for morality, and that egoistic motivation is incompatible with it, is shared by many authors (see Stich et al., 2010, p. 148; see also Kraut, 2020). For example, Shafer-Landau (2012) claims that "[i]f psychological egoism is true, then we can't be altruistic", and, since we cannot be required to do the impossible, "[i]f we can't be altruistic, then it can't be our duty to be altruistic" (p. 93). The truth of psychological altruism is considered by authors like him as an important condition for morality as we know it. Shafer-Landau (2012) claims that, if psychological *egoism* is true, then "[w]e would have to radically change our moral ideals, ridding them of altruistic elements" (p. 93). However, if my analysis in this thesis is correct, it is hard to justify these claims.

The truth or falsehood of psychological altruism does not have any clear relevant consequences, neither to individuals' behaviors nor to their subjective states[101]. If true, psychological altruism does not imply that people are more likely to help others, since, as discussed in Chapter 5, egoistic motivation can easily be structured so as to produce the same behavioral outcomes as altruistic motivation. Once we define altruism as an ultimate desire to increase others' welfare, we see that it may be unconscious, may never produce helping behaviors, and may be present even in the most selfish individuals. So, if we want to attribute moral relevance to altruistic motivation, we should not define it in terms of ultimate desires to increase the welfare of others. Virtue altruism offers an alternative way of conceiving altruism, which preserves its moral relevance.

[101] There are consequences, such as the fact that people who believe in psychological altruism are more likely to act altruistically than people who believe in psychological egoism, as discussed in the first chapter. But this is an effect produced by the *belief* in the hypotheses, not by the phenomena they describe.

The opposition between psychological altruism and psychological egoism, at first, seems to be a fundamental debate about human nature. When we hear about the altruism question, wondering whether humans can be altruists or are condemned to be egoists, we are left with the impression that this question touches something very deep about who we are as human beings. However, once we have a deeper understanding of the standard account of altruistic motivation, the importance of the altruism question is no longer obvious. If there is a relevant issue about human nature lurking underneath the question of whether we can have altruistic motivation, this is left behind in the contemporary overly abstract way of thinking about altruistic motivation as an ultimate desire to increase the welfare of others.

Batson (2011) complained that "most of the rather vast literatures in biology, primatology, behavioral economics, and developmental and social psychology… that claims to provide data on altruism… rarely even aspire to address the motivational issues" (p. 29). He mentions this to highlight what he considers to be a flagrant flaw in the literature. But after the problems raised in this thesis, we can see that there are actually good reasons for researchers to not engage with the altruism question in the terms proposed by Batson and the contemporary philosophical literature.

The main goal of this thesis was to present a multifaceted criticism of the standard account of altruistic motivation. In doing so, I explored many different aspects of altruism. I hope that my work, on top of helping researchers to avoid the intrinsically problematic standard account of altruistic motivation, can also inspire future research on altruism. Beyond the philosophical, abstract issues involved in the discussion about altruism, the concepts we use to characterize altruism and egoism can have many practical effects. Researchers and activists should have better ways of thinking about altruism. For example, I believe that the theoretical

framework of movements such as the Effective Altruism movement (see Singer, 2015) could be benefited from a richer discussion regarding the conceptualizations of altruistic motivation. Philosophical work on the complex and fascinating idea of altruism can influence the way we think about our concern for others. I hope that my criticism of the standard account of altruistic motivation may encourage new forms of thinking about (and practicing) altruism.

References

Abbot, P., Abe, J., Alcock, J., Alizon, S., Alpedrinha, J. A. C., Andersson, M., Andre, J.-B., van Baalen, M., Balloux, F., Balshine, S., Barton, N., Beukeboom, L. W., Biernaskie, J. M., Bilde, T., Borgia, G., Breed, M., Brown, S., Bshary, R., Buckling, A., … Zink, A. (2011). Inclusive fitness theory and eusociality. *Nature, 471*(7339), E1–E4. https://doi.org/10.1038/nature09831

Alexandra, A. (1992). Should Hobbes's state of nature be represented as a prisoner's dilemma? *The Southern Journal of Philosophy, 30*(2), 1–16. https://doi.org/10.1111/j.2041-6962.1992.tb01712.x

Allen, C., & Bekoff, M. (1999). *Species of mind: The philosophy and biology of cognitive ethology*. MIT Press.

Alvarez, M. (2009). How many kinds of reasons? *Philosophical Explorations, 12*(2), 181–193. https://doi.org/10.1080/13869790902838514

Aristotle. (2000). *Nicomachean ethics* (R. Crisp, Ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511802058

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition, 13*(3), 263–308. https://doi.org/10.1016/0010-0277(83)90012-4

Arpaly, N., & Schroeder, T. (2014). *In praise of desire*. Oxford University Press.

Ayer, A. J. (1971). *Language, truth, and logic*. Penguin Books.

Barragan, R. C., & Dweck, C. S. (2014). Rethinking natural altruism: Simple reciprocal interactions trigger children's benevolence. *Proceedings of the National Academy of Sciences, 111*(48), 17071–17074. https://doi.org/10.1073/pnas.1419408111

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 629–654. https://doi.org/10.1037/0278-7393.11.1-4.629

Batson, C. D. (1987). Prosocial motivation: Is it ever truly altruistic? In *Advances in experimental social psychology* (Vol. 20, pp. 65–122). Elsevier. https://doi.org/10.1016/S0065-2601(08)60412-8

Batson, C. D. (1991). *The altruism question: Toward a social psychological answer*. Lawerence Erlbaum.

Batson, C. D. (2000). Unto others: A service... and a disservice. *Journal of Consciousness Studies*, *7*(1–2), 207–210.

Batson, C. D. (2011). *Altruism in humans*. Oxford University Press.

Batson, C. D. (2018). *A scientific search for altruism: Do we only care about ourselves?* Oxford University Press.

Batson, C. D., Duncan, B. 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. https://doi.org/10.1037/0022-3514.40.2.290

Batson, C. D., Early, S., & Salvarani, G. (1997). Perspective taking: Imagining how another feels versus imaging how you would feel. *Personality and Social Psychology Bulletin*, *23*(7), 751–758. https://doi.org/10.1177/0146167297237008

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

Batson, C. D., Turk, C. L., Shaw, L. L., & Klein, T. R. (1995). Information function of empathic emotion: Learning that we value the other's welfare. *Journal of Personality and Social Psychology*, *68*, 300–313. https://doi.org/10.1037/0022-3514.68.2.300

Bennett, K. (2007). Mental causation. *Philosophy Compass*, *2*(2), 316–337. https://doi.org/10.1111/j.1747-9991.2007.00063.x

Bentham, J. (1843). *The works of Jeremy Bentham, published under the superintendence of his executor, John Bowring* (Vol. 9). William Tait.

Bentham, J. (2000). *An introduction to the principles of morals and legislation*. Batoche Books. (Original work published 1789)

Berman, S. (2003). A Defense of Psychological Egoism. In N. Reshotko (Ed.), *Desire, Identity and Existence*. Academic Printing and Publishing.

Blackburn, S. (1998). *Ruling passions: A theory of practical reasoning*. Oxford University Press.

Blakemore, R. P., & Frankel, R. B. (1981). Magnetic navigation in bacteria. *Scientific American*, *245*(6), 58–65.

Bloom, P. (2018). *Against empathy: The case for rational compassion*. Ecco.

Borrello, M. E. (2010). *Evolutionary restraints: The contentious history of group selection*. University of Chicago Press.

Bourdeau, M. (2022). Auguste Comte. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2022/entries/comte/

Broadie, A., & Smith, C. (2022). Scottish philosophy in the 18th century. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2022/entries/scottish-18th/

Brooks, E. B., & Brooks, A. T. (1998). *The original Analects: Sayings of Confucius and his successors*. Columbia University Press.

Buller, D. J. (2005). *Adapting minds: Evolutionary psychology and the persistent quest for human nature*. MIT Press.

Burge, T. (1992). Philosophy of language and mind, 1950–1990. *Philosophical Review*, *101*(1), 3–51.

Butler, J. (2006). Fifteen sermons preached at the Rolls Chapel. In D. E. White (Ed.), *The works of Bishop Butler* (pp. 33–147). University of Rochester Press. (Original work published 1726)

Cambridge Dictionary. (n.d.). Altruism. In *Cambridge Dictionary*. Retrieved October 13, 2022, from https://dictionary.cambridge.org/dictionary/english/altruism

Carey, S., & Spelke, E. (1996). Science and core knowledge. *Philosophy of Science*, *63*(4), 515–533. https://doi.org/10.1086/289971

Carter, A. (2005). Evolution and the problem of altruism. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *123*(3), 213–230. https://doi.org/10.1007/s11098-005-1289-6

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, *78*(2), 67–90. https://doi.org/10.2307/2025900

Churchland, P. M. (1985). Reduction, qualia, and the direct introspection of brain states. *The Journal of Philosophy*, *82*(1), 8. https://doi.org/10.2307/2026509

Churchland, P. S. (1989). *Neurophilosophy: Toward a unified science of the mind-brain*. MIT Press.

Cialdini, R. B., Baumann, D. J., & Kenrick, D. T. (1981). Insights from sadness: A three-step model of the development of altruism as hedonism. *Developmental Review*, *1*(3), 207–223. https://doi.org/10.1016/0273-2297(81)90018-6

Cialdini, R. B., & Kenrick, D. T. (1976). Altruism as hedonism: A social development perspective on the relationship of negative mood state and helping. *Journal of Personality and Social Psychology*, *34*, 907–914. https://doi.org/10.1037/0022-3514.34.5.907

Clavien, C. (2012). Altruistic emotional motivation: An argument in favour of psychological altruism. In K. S. Plaisance & T. A. C. Reydon (Eds.), *Philosophy of behavioral biology* (Vol. 282, pp. 275–296). Springer Netherlands. https://doi.org/10.1007/978-94-007-1951-4_13

Clavien, C., & Chapuisat, M. (2013). Altruism across disciplines: One word, multiple meanings. *Biology & Philosophy*, *28*(1), 125–140. https://doi.org/10.1007/s10539-012-9317-3

Clavien, C., & Chapuisat, M. (2016). The evolution of utility functions and psychological altruism. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *56*, 24–31. https://doi.org/10.1016/j.shpsc.2015.10.008

Clavien, C., & Klein, R. A. (2010). Eager for fairness or for revenge? Psychological altruism in economics. *Economics and Philosophy*, *26*(3), 267–290. https://doi.org/10.1017/S0266267110000374

Comte, A. (1875-1877). *System of positive polity* (J. H. Bridges, Trans.; Vol. 1–4). Longmans, Green and Company. (Original work published 1851–1854)

Comte, A. (2000). *The positive philosophy of Auguste Comte* (H. Martineau, Trans.; Vols. 1–3). Batoche Books. (Original work published 1853)

Comte, A. (2009). *The Catechism of Positive Religion: Or summary exposition of the universal religion in thirteen systematic conversations between a woman and a priest of Humanity*. Cambridge University Press. (Original work published 1852)

Corns, J. (2016). Pain eliminativism: Scientific and traditional. *Synthese*, *193*(9), 2949–2971. https://doi.org/10.1007/s11229-015-0897-8

Coulter, I. D., Wilkes, M., & Der-Martirosian, C. (2007). Altruism revisited: A comparison of medical, law and business students' altruistic attitudes. *Medical Education*, *41*(4), 341–345. https://doi.org/10.1111/j.1365-2929.2007.02716.x

Crimmins, J. E. (2021). Jeremy Bentham. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/bentham/

Crisp, R. (2019). *Sacrifice regained: Morality and self-interest in British moral philosophy from Hobbes to Bentham*. Oxford University Press.

Csikszentmihalyi, M. (2020). Confucius. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2020/entries/confucius/

Darwall, S. L. (1974). Nagel's argument for altruism. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *25*(2), 125–130.

Darwall, S. L. (1995). *The British moralists and the internal "ought" 1640-1740*. Cambridge University Press.

Darwin, C. (2008). *On the origin of species* (G. Beer, Ed.; Rev. ed). Oxford University Press. (Original work published 1859)

Darwin, C. (2009). *The descent of man and selection in relation to sex* (Vol. 1). Cambridge University Press. (Original work published 1871)

Dawkins, R. (2006). *The selfish gene* (30th anniversary ed). Oxford University Press. (Original work published 1976)

de Waal, F. B. M. (2006). *Primates and philosophers: How morality evolved*. Princeton University Press.

de Waal, F. B. M. (2008). Putting the altruism back into altruism: The evolution of empathy. *Annual Review of Psychology*, *59*(1), 279–300. https://doi.org/10.1146/annurev.psych.59.103006.093625

de Waal, F. B. M. (2012). The antiquity of empathy. *Science*, *336*(6083), 874–876. https://doi.org/10.1126/science.1220999

Dennett, D. C. (1978). Why you can't make a computer that feels pain. *Synthese*, *38*(3), 415–456.

Dennett, D. C. (1989). *The intentional stance*. MIT Press.

Descartes, R. (1985). Meditations on first philosophy. In J. Cottingham, R. Stoothoff, & D. Murdoch (Trans.), *The philosophical writings of Descartes* (pp. 1–62). Cambridge University Press. (Original work published 1641)

Dickinson, A., & Balleine, B. (2010). Hedonics: The cognitive-motivational interface. In M. L. Kringelbach & K. C. Berridge (Eds.), *Pleasures of the brain* (pp. 74–84). Oxford University Press.

Dixon, T. (2008). *The invention of altruism: Making moral meanings in Victorian Britain*. British Academy. https://doi.org/10.5871/bacad/9780197264263.001.0001

Doris, J., Stich, S., & Walmsley, L. (2020). Empirical approaches to altruism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/altruism-empirical/

Dretske, F. I. (1991). *Explaining behavior: Reasons in a world of causes*. MIT Press.

Dubs, H. H. (1951). The development of altruism in Confucianism. *Philosophy East and West*, *1*(1), 48. https://doi.org/10.2307/1396935

Durant, W. (1962). *Story of philosophy*. Time Incorporated.

Eisen, S. (1967). Herbert Spencer and the Spectre of Comte. *Journal of British Studies*, *7*(1), 48–67.

Eisenberg, N., & Miller, P. A. (1987). The relation of empathy to prosocial and related behaviors. *Psychological Bulletin*, *101*(1), 91–119. https://doi.org/10.1037/0033-2909.101.1.91

Eisenberg, N., Schaller, M., Fabes, R. A., Bustamante, D., Mathy, R. M., Shell, R., & Rhodes, K. (1988). Differentiation of personal distress and sympathy in children and adults. *Developmental Psychology*, *24*, 766–775. https://doi.org/10.1037/0012-1649.24.6.766

English Standard Version Bible. (2001). ESV Online. https://esv.literalword.com/

Ewin, R. E. (1991). *Virtues and rights: The moral philosophy of Thomas Hobbes*. Westview Press.

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185–190. https://doi.org/10.1016/j.tics.2004.02.007

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140. https://doi.org/10.1038/415137a

Feigin, S., Owens, G., & Goodyear-Smith, F. (2014). Theories of human altruism: A systematic review. *Journal of Psychiatry and Brain Functions*, *1*(1), 5. https://doi.org/10.7243/2055-3447-1-5

Feigl, H. (1958). The ‘mental’ and the ‘physical’. *Minnesota Studies in the Philosophy of Science*, *2*(2), 370–497.

Feinberg, J. (2013). Psychological egoism. In J. Feinberg & R. Shafer-Landau (Eds.), *Reason and responsibility: Readings in some basic problems of philosophy* (pp. 501–513). Cengage Learning.

Fodor, J. A. (1981). *Representations: Philosophical essays on the foundations of cognitive science*. The Harvester Press.

Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions*. W. W. Norton & Company.

Frank, R. H., Gilovich, T., & Regan, D. T. (1993). Does studying economics inhibit cooperation? *Journal of Economic Perspectives*, *7*(2), 159–171. https://doi.org/10.1257/jep.7.2.159

Frankel, R. B., & Bazylinski, D. A. (2002). Magnetotaxis: Microbial. In *Encyclopedia of life sciences*. John Wiley & Sons, Ltd. https://doi.org/10.1038/npg.els.0000397

Frankel, R. B., & Blakemore, R. P. (1989). Magnetite and magnetotaxis in microorganisms. *Bioelectromagnetics*, *10*(3), 223–237. https://doi.org/10.1002/bem.2250100303

Fraser, C. (2022). Mohism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2022/entries/mohism/

Galston, W. A. (1993). Cosmopolitan altruism. *Social Philosophy and Policy*, *10*(1), 118–134. https://doi.org/10.1017/S0265052500004040

Garson, J. (2015). *The biological mind: A philosophical introduction*. Routledge.

Garson, J. (2016). Two types of psychological hedonism. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *56*, 7–14. https://doi.org/10.1016/j.shpsc.2015.10.011

Gaus, G. F. (1990). *Value and justification: The foundations of liberal theory*. Cambridge University Press.

Gert, B. (1967). Hobbes and psychological egoism. *Journal of the History of Ideas*, *28*(4), 503–520. https://doi.org/10.2307/2708526

Gill, M. B. (2021). Lord Shaftesbury [Anthony Ashley Cooper, 3rd Earl of Shaftesbury]. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/shaftesbury/

Goldman, A. I. (1970). *A theory of human action*. Prentice-Hall.

Graham, G. (2019). Behaviorism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2019). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2019/entries/behaviorism/>

Gregory, A. (2021). *Desire as belief: A study of desire, motivation, and rationality*. Oxford University Press. https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198848172.001.0001/oso-9780198848172

Grote, S. (2010). Shaftesbury's egoistic hedonism. *Aufklärung*, *22*, 135–149.

Guillin, V. (2018). Comte and social science. In M. Bourdeau, M. Pickering, & W. Schmaus (Eds.), *Love, order, and progress: The science, philosophy, and politics of Auguste Comte* (pp. 128–160). University of Pittsburgh Press.

Hacking, I. (1995). *Rewriting the soul: Multiple personality and the sciences of memory*. Princeton University Press.

Hamilton, W. D. (1964a). The genetical evolution of social behaviour I. *Journal of Theoretical Biology*, *7*(1), 1–16. https://doi.org/10.1016/0022-5193(64)90038-4

Hamilton, W. D. (1964b). The genetical evolution of social behaviour II. *Journal of Theoretical Biology*, *7*(1), 17–52. https://doi.org/10.1016/0022-5193(64)90039-6

Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development*, *26*(1), 30–39. https://doi.org/10.1016/j.cogdev.2010.09.001

Hampton, J. (1995). *Hobbes and the social contract tradition*. Cambridge University Press.

Hampton, J. (1999). Concepts. In R. A. Wilson & F. C. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences* (pp. 176–179). MIT Press.

Hardcastle, V. G. (1997). When a pain is not. *The Journal of Philosophy*, *94*(8), 381. https://doi.org/10.2307/2564606

Harman, G. (2000). Can evolutionary theory provide evidence against psychological hedonism? *Journal of Consciousness Studies*, *7*(1–2), 219–221.

Harman, O. (2014). A history of the altruism–morality debate in biology. *Behaviour*, *151*(2–3), 147–165. https://doi.org/10.1163/1568539X-00003133

Hausman, D. M. (2012). *Preference, value, choice, and welfare*. Cambridge University Press.

Hempel, C. (1959). The empiricist criterion of meaning. In A. J. Ayer (Ed.), *Logical Positivism* (pp. 53–59). The Free Press. (Original work published 1930)

Hepach, R., Vaish, A., & Tomasello, M. (2013). A new look at children's prosocial motivation: Children's prosocial motivation. *Infancy*, *18*(1), 67–90. https://doi.org/10.1111/j.1532-7078.2012.00130.x

Hobbes, T. (1928). *Elements of law, natural and politic*. Cambridge University Press. (Original work published 1640)

Hobbes, T. (1983). *De cive: The English version entitled, in the first edition, philosophical rudiments concerning government and society* (H. Warrender, Ed.). Clarendon Press. (Original work published 1642)

Hobbes, T. (1998). *Leviathan* (J. C. A. Gaskin, Ed.). Oxford University Press. (Original work published 1651)

Hoffman, M. L. (1996). Empathy and moral development. *The Annual Report of Educational Psychology in Japan*, *35*(0), 157–162. https://doi.org/10.5926/arepj1962.35.0_157

Hume, D. (1960). *A treatise of human nature* (L. A. Selby-Bigge, Ed.). Clarendon Press. (Original work published 1739)

Hursthouse, R., & Pettigrove, G. (2018). Virtue ethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2018). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/

Hutcheson, F. (2002). *An essay on the nature and conduct of the passions and affections, with illustrations on the moral sense* (A. Garrett, Ed.). Liberty Fund. (Original work published 1728)

Hutcheson, F. (2004). *An inquiry into the original of our ideas of beauty and virtue in two treatises* (W. Leidhold, Ed.). Liberty Fund. (Original work published 1725)

Hutto, D., & Ravenscroft, I. (2021). Folk psychology as a theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/folkpsych-theory/

Huxley, T. H. (1902). The struggle for existence in human society. In *Evolution and ethics: And other essays* (Vol. 9, pp. 195–236). D. Appleton. (Original work published 1888)

Ingold, T. (2016). What is a social relationship. In T. Ingold (Ed.), *Evolution and social life*. Routledge. (Original work published 1986)

Jackson, F. (1982). Epiphenomenal qualia. *The Philosophical Quarterly*, *32*(127), 127. https://doi.org/10.2307/2960077

Jamieson, D. (2002). Sober and Wilson on psychological altruism. *Philosophy and Phenomenological Research*, *65*(3), 702–710. https://doi.org/10.1111/j.1933-1592.2002.tb00236.x

Johnsrude, I. S., Owen, A. M., Zhao, W. V., & White, N. M. (1999). Conditioned preference in humans: A novel experimental approach. *Learning and Motivation*, *30*(3), 250–264. https://doi.org/10.1006/lmot.1999.1031

Kahn, C. H. (1981). Aristotle and altruism. *Mind, New Series*, *90*(357), 20–40.

Kahn, C. H. (1987). Plato's theory of desire. *The Review of Metaphysics*, *41*(1), 77–103.

Kaiser, K. (2017). A new taxonomy of altruism in terms of prosocial behaviors. *Dialogue & Nexus*, *4*, 8.

Kauppinen, A. (2022). Moral Sentimentalism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2022/entries/moral-sentimentalism/

Kavka, G. S. (1983). Hobbes's war of all against all. *Ethics*, *93*(2), 291–310.

Kavka, G. S. (1986). *Hobbesian moral and political theory*. Princeton University Press.

Kidson, S. H., & Fabian, B. C. (1981). The effect of temperature on tyrosinase activity in Himalayan mouse skin. *Journal of Experimental Zoology*, *215*(1), 91–97. https://doi.org/10.1002/jez.1402150111

Kitcher, P. (1993). The evolution of human altruism. *The Journal of Philosophy*, *90*(10), 497. https://doi.org/10.2307/2941024

Kitcher, P. (1998). Psychological altruism, evolutionary origins, and moral rules. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *89*(2), 283–316.

Kitcher, P. (2010). Varieties of altruism. *Economics and Philosophy*, *26*(2), 121–148. https://doi.org/10.1017/S0266267110000167

Kitcher, P. (2011). *The ethical project*. Harvard University Press.

Kraut, R. (2020). Altruism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/altruism/

Kropotkin, P. (2021). *Mutual aid: A factor of evolution*. Black Rose Books Ltd. (Original work published 1902)

Lakoff, G. (2007). Cognitive models and prototype theory. In V. Evans, B. K. Bergen, & J. Zinken (Eds.), *The cognitive linguistics reader*. Equinox.

LaFollette, H. (1988). The truth in psychological egoism. In J. Feinberg (Ed.), *Reason and responsibility* (pp. 500–507). Wadsworth.

Lefevre, C. T., & Bazylinski, D. A. (2013). Ecology, diversity, and evolution of magnetotactic bacteria. *Microbiology and Molecular Biology Reviews*, *77*(3), 497–526. https://doi.org/10.1128/MMBR.00021-13

Lemos, J. (2004). Psychological hedonism, evolutionary biology, and the experience machine. *Philosophy of the Social Sciences*, *34*(4), 506–526. https://doi.org/10.1177/0048393104269597

Lemos, J. (2008). *Commonsense Darwinism: Evolution, morality, and the human condition*. Open Court.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, *50*(3), 249–258. https://doi.org/10.1080/00048407212341301

Lewis, D. (1988). Desire as belief. *Mind*, *97*(387), 323–332.

Lewis, D. (1996). Desire as belief II. *Mind*, *105*(418), 303–313.

Lewontin, R. C. (1970). The units of selection. *Annual Review of Ecology and Systematics*, *1*(1), 1–18.

Lichtenberg, J. (2008). About altruism. *Philosophy and Public Policy Quarterly*, *28*(1), 2–6.

Lloyd, S. A., & Sreedhar, S. (2022). Hobbes's moral and political philosophy. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/hobbes-moral/

Ludwig, K. (2003). The mind–body problem: An overview. In S. P. Stich & T. A. Warfield (Eds.), *The Blackwell guide to philosophy of mind* (pp. 1–46). Blackwell Publishing Ltd.

Lupyan, G. (2013). The difficulties of executing simple algorithms: Why brains make mistakes computers don't. *Cognition, 129*(3), 615–636. https://doi.org/10.1016/j.cognition.2013.08.015

Lycan, W. G. (2003). The mind–body problem. In S. P. Stich & T. A. Warfield (Eds.), *The Blackwell guide to philosophy of mind* (pp. 47–64). Blackwell Publishing Ltd.

Lycan, W. G. (2012). Desire considered as a propositional attitude. *Philosophical Perspectives*, *26*(1), 201–215. https://doi.org/10.1111/phpe.12003

MacIntyre, A. (1967). Egoism and altruism. In P. Edwards (Ed.), *The encyclopedia of philosophy* (Vol. 2, pp. 462–466). Macmillan.

Maibom, H. L. (2009). Feeling for others: Empathy, sympathy, and morality. *Inquiry*, *52*(5), 483–499. https://doi.org/10.1080/00201740903302626

Maibom, H. L. (2012). The many faces of empathy and their relation to prosocial action and aggression inhibition. *WIREs Cognitive Science*, *3*(2), 253–263. https://doi.org/10.1002/wcs.1165

Mandeville, B. (1988). *The fable of the bees or private vices, publick benefits* (Vol. 1–2). Liberty Fund. (Original work published 1714)

Martin, G. B., & Clark, R. D. (1982). Distress crying in neonates: Species and peer specificity. *Developmental Psychology*, *18*, 3–9. https://doi.org/10.1037/0012-1649.18.1.3

Mason, E. (2018). Value pluralism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2018). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2018/entries/value-pluralism/

Maurer, C. (2019). *Self-love, egoism and the selfish hypothesis: Key debates from eighteenth-century British moral philosophy*. Edinburgh University Press.

McNeilly, F. S. (1966). Egoism in Hobbes. *The Philosophical Quarterly*, *16*(64), 193. https://doi.org/10.2307/2218463

McVeigh, R. (2020). The neurosociology of Auguste Comte. *Social Science Information*, *59*(2), 329–354. https://doi.org/10.1177/0539018420922759

Merriam-Webster. (n.d.). Altruism. In *Merriam-Webster.com dictionary*. Retrieved October 13, 2022, from https://www.merriam-webster.com/dictionary/altruism

Mill, J. S. (1969). Auguste Comte and positivism. In J. M. Robson (Ed.), *Essays on ethics, religion, and society*. University of Toronto Press. (Original work published 1865)

Moore, A. (2019). Hedonism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2019). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/hedonism/

Moorlock, G., Ives, J., & Draper, H. (2014). Altruism in organ donation: An unnecessary requirement? *Journal of Medical Ethics*, *40*(2), 134–138. https://doi.org/10.1136/medethics-2012-100528

Morillo, C. R. (1990). The reward event and motivation. *The Journal of Philosophy*, *87*(4), 169. https://doi.org/10.2307/2026679

Nagel, T. (1970). *The possibility of altruism*. Princeton University Press.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435–450. https://doi.org/10.2307/2183914

Newey, G. (2008). *Routledge philosophy guidebook to Hobbes and Leviathan*. Routledge.

Nietzsche, F. (1998). *On the Genealogy of Morality*. Hackett Publishing. (Original work published 1887)

Nowak, M. A., Tarnita, C. E., & Wilson, E. O. (2010). The evolution of eusociality. *Nature*, *466*(7310), 1057–1062. https://doi.org/10.1038/nature09205

Nuffield Council on Bioethics. (2011). *Human bodies: Donation for medicine and research*. Nuffield Council on Bioethics. https://www.nuffieldbioethics.org/publications/human-bodies-donation-for-medicine-and-research

Okasha, S. (2018). *Agents and goals in evolution*. Oxford University Press.

Oliner, S. P., & Oliner, P. M. (1988). *The altruistic personality: Rescuers of Jews in Nazi Europe*. The Free Press.

Oswald, P. A. (1996). The effects of cognitive and affective perspective taking on empathic concern and altruistic helping. *The Journal of Social Psychology*, *136*(5), 613–623. https://doi.org/10.1080/00224545.1996.9714045

Peacock, M. S., Schefczyk, M., & Schaber, P. (2005). Altruism and the indispensability of motives. *Analyse & Kritik*, *27*(1), 188–196. https://doi.org/10.1515/auk-2005-0111

Peart, S., & Levy, D. M. (2005). *The "vanity of the philosopher": From equality to hierarchy in postclassical economics*. University of Michigan Press.

Penner, L. A., Fritzsche, B. A., Craiger, J. P., & Freifeld, T. S. (2014). Measuring the prosocial personality. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment* (Vol. 10). Psychology Press.

Pennings, G. (2015). Central role of altruism in the recruitment of gamete donors. *Monash Bioethics Review*, *33*(1), 78–88. https://doi.org/10.1007/s40592-015-0019-x

Piccinini, G., & Schulz, A. W. (2018). The evolution of psychological altruism. *Philosophy of Science*, *85*(5), 1054–1064. https://doi.org/10.1086/699743

Piccinini, G., & Schulz, A. W. (2019). The ways of altruism. *Evolutionary Psychological Science*, *5*(1), 58–70. https://doi.org/10.1007/s40806-018-0167-3

Place, U. T. (1956). Is consciousness a brain process? *British Journal of Psychology*, *47*(1), 44–50.

Plato. (1997). *Complete works* (J. M. Cooper & D. S. Hutchinson, Eds.). Hackett Publishing Company.

Pollock, J. L. (2006). *Thinking about acting: Logical foundations for rational decision making*. Oxford University Press.

Post, S. G. (2005). Altruism, happiness, and health: It's good to be good. In G. Ironson & L. Powell (Eds.), *An exploration of the health benefits of factors that help us to thrive*. Psychology Press.

Putnam, H. (1967). The nature of mental states. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 51–58). Pittsburgh University Press.

Putnam, H. (1975). Philosophy and our mental life. In *Philosophical papers* (Vol. 2, pp. 291–303). Cambridge University Press. https://doi.org/10.1017/CBO9780511625251.016

Rachels, S., & Rachels, J. (2015). *The elements of moral philosophy* (8th ed.). McGraw-Hill Education.

Ramsey, G. (2016). Can altruism be unified? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *56*, 32–38. https://doi.org/10.1016/j.shpsc.2015.10.007

Rizzolatti, G. (2005). The mirror neuron system and its function in humans. *Anatomy and Embryology*, *210*(5–6), 419–421. https://doi.org/10.1007/s00429-005-0039-z

Rokeach, M. (1973). *The nature of human values* (pp. x, 438). Free Press.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328–350. https://doi.org/10.1016/0010-0285(73)90017-0

Rosch, E. H. (1978). Principles of categorization. In B. Lloyd & E. Rosch (Eds.), *Cognition and categorization*. Lawrence Erlbaum.

Rottschaefer, W. A. (2000). It's been a pleasure, but that's not why I did it: Are Sober and Wilson too generous toward their selfish brethren? *Journal of Consciousness Studies*, *7*(1–2), 239–243.

Ryle, G. (2009). *The concept of mind*. Routledge. (Original work published 1949)

Schefczyk, M., & Peacock, M. (2010). Altruism as a thick concept. *Economics and Philosophy*, *26*(2), 165–187. https://doi.org/10.1017/S0266267110000180

Schlick, M. (1959). The turning point in philosophy. In A. J. Ayer (Ed.), *Logical positivism* (pp. 108–132). The Free Press. (Original work published 1950)

Schmitter, A. M. (2021). 17th and 18th century theories of emotions. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/emotions-17th18th/

Schneider, S. M., & Morris, E. K. (1987). A history of the term radical behaviorism: From Watson to Skinner. *The Behavior Analyst*, *10*(1), 27–39.

Schroeder, M. (2021). Value theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/value-theory/

Schroeder, T. (2004). *Three faces of desire*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195172379.001.0001

Schroeder, T. (2020). Desire. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2020/entries/desire/

Schulz, A. W. (2011). Sober & Wilson's evolutionary arguments for psychological altruism: A reassessment. *Biology & Philosophy*, *26*(2), 251–260. https://doi.org/10.1007/s10539-009-9179-5

Schulz, A. W. (2016). Altruism, egoism, or neither: A cognitive-efficiency-based evolutionary biological perspective on helping behavior. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *56*, 15–23. https://doi.org/10.1016/j.shpsc.2015.10.006

Schulz, A. W. (2018). *Efficient cognition: The evolution of representational decision making*. MIT Press.

Searle, J. R. (1983). *Intentionality, an essay in the philosophy of mind*. Cambridge University Press.

Sellars, W. (1956). Empiricism and the philosophy of mind. *Minnesota Studies in the Philosophy of Science, 1*(19), 253–329.

Seyfarth, R. M., & Cheney, D. L. (2012). The evolutionary origins of friendship. *Annual Review of Psychology*, *63*(1), 153–177. https://doi.org/10.1146/annurev-psych-120710-100337

Shafer-Landau, R. (2012). *The fundamentals of ethics* (2nd ed). Oxford University Press.

Shaftesbury, A. A. C. (2001). *Characteristicks of men, manners, opinions, times* (D. den Uyl, Ed.; Vol. 1–3). Liberty Fund. (Original work published 1711)

Sheridan, P. (2007). Parental affection and self-interest: Mandeville, Hutcheson, and the question of natural benevolence. *History of Philosophy Quarterly*, *24*(4), 377–392.

Simner, M. L. (1971). Newborn's response to the cry of another infant. *Developmental Psychology*, *5*(1), 136–150. https://doi.org/10.1037/h0031066

Singer, P. (2015). *The most good you can do: How effective altruism is changing ideas about living ethically*. Yale University Press.

Skyrms, B. (2014). *Evolution of the social contract* (2nd ed.). Cambridge University Press.

Smart, J. J. C. (1959). Sensations and brain processes. *The Philosophical Review*, *68*(2), 141. https://doi.org/10.2307/2182164

Smith, A. (2002). *The theory of moral sentiments* (K. Haakonssen, Ed.). Cambridge University Press. (Original work published 1759)

Smith, M. (1994). *The moral problem*. Blackwell.

Sober, E. (1988). What is evolutionary altruism? *Canadian Journal of Philosophy Supplementary Volume, 14*, 75–99. https://doi.org/10.1080/00455091.1988.10715945

Sober, E. (1992). Hedonism and Butler's stone. *Ethics*, *103*(1), 97–103. https://doi.org/10.1086/293472

Sober, E. (1994). *From a biological point of view: Essays in evolutionary philosophy*. Cambridge University Press.

Sober, E. (2001). The two faces of fitness. In S. Rama, P. Diane, C. Krimbas, & J. Beatty (Eds.), *Thinking about evolution: Historical, philosophical and political perspectives* (Vol. 2, pp. 25–38). Cambridge University Press.

Sober, E. (2013). Psychological egoism. In H. LaFollette & I. Persson (Eds.), *The Blackwell guide to ethical theory* (2nd ed., pp. 148–168). John Wiley & Sons.

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

Sober, E., & Wilson, D. S. (2000). Morality and 'Unto Others'. Response to commentary discussion. *Journal of Consciousness Studies, 7*(1–2), 257–268.

Spencer, H. (1883). The data of ethics. In *The principles of ethics* (pp. 1–288). D. Appleton and Company. (Original work published 1862)

Sterelny, K. (2003). *Thought in a hostile world: The evolution of human cognition*. Blackwell Publishing.

Stich, S. (1978). Beliefs and subdoxastic states. *Philosophy of Science, 45*(4), 499–518.

Stich, S. (2007). Evolution, altruism and cognitive architecture: A critique of Sober and Wilson's argument for psychological altruism. *Biology & Philosophy, 22*(2), 267–281. https://doi.org/10.1007/s10539-006-9030-1

Stich, S. (2016). Why there might not be an evolutionary explanation for psychological altruism. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 56*, 3–6. https://doi.org/10.1016/j.shpsc.2015.10.005

Stich, S., Doris, J. M., & Cushman, F. (2010). Altruism. In J. M. Doris & Moral Psychology Research Group (Eds.), *The moral psychology handbook*. Oxford University Press.

Stich, S., & Nichols, S. (2003). Folk psychology. In S. P. Stich & T. A. Warfield (Eds.), *The Blackwell guide to philosophy of mind* (pp. 235–255). Blackwell Publishing Ltd.

Strawson, G. (2010). *Mental reality* (2nd ed.). MIT Press.

Stueber, K. (2019). Empathy. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2019). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2019/entries/empathy/

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Titchener, E. B. (2014). Introspection and empathy. *Dialogues in Philosophy, Mental & Neuro Sciences, 7*(1). (Original work published 1909)

Todes, D. P. (1989). *Darwin without Malthus: The struggle for existence in Russian evolutionary thought*. Oxford University Press.

Trivers, R. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35–57. https://doi.org/10.1086/406755

Trivers, R. (1972). Parental investment and sexual selection. In B. G. Campbell (Ed.), *Sexual selection and the descent of man: The Darwinian pivot*. Aldine Transaction.

van Rysewyk, S. (2016). Is pain unreal? In S. van Rysewyk (Ed.), *Meanings of pain* (pp. 71–86). Springer International Publishing. https://doi.org/10.1007/978-3-319-49022-9_5

Wallach, L., & Wallach, M. A. (1991). Why altruism, even though it exists, cannot be demonstrated by social psychological experiments. *Psychological Inquiry*, *2*(2), 153–155. https://doi.org/10.1207/s15327965pli0202_15

Wang, L., Malhotra, D., & Murnighan, J. K. (2011). Economics education and greed. *Academy of Management Learning & Education*, *10*(4), 643–660. https://doi.org/10.5465/amle.2009.0185

Warneken, F., Hare, B., Melis, A. P., Hanus, D., & Tomasello, M. (2007). Spontaneous altruism by chimpanzees and young children. *PLoS Biology*, *5*(7), e184. https://doi.org/10.1371/journal.pbio.0050184

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, *311*(5765), 1301–1303. https://doi.org/10.1126/science.1121448

Warneken, F., & Tomasello, M. (2008). Extrinsic rewards undermine altruistic tendencies in 20-month-olds. *Developmental Psychology*, *44*(6), 1785–1788. https://doi.org/10.1037/a0013860

Weinstein, D. (2019). Herbert Spencer. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2019). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2019/entries/spencer/

Welchman, J. (2007). Who rebutted Bernard Mandeville? *History of Philosophy Quarterly*, *24*(1), 57–74.

Wernick, A. (2001). *Auguste Comte and the Religion of Humanity: The post-theistic program of French social theory*. Cambridge University Press.

West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology*, *20*(2), 415–432. https://doi.org/10.1111/j.1420-9101.2006.01258.x

Wickens, A. P. (2014). *A history of the brain*. Psychology Press. https://doi.org/10.4324/9781315794549

Wilkinson, G. S. (1984). Reciprocal food sharing in the vampire bat. *Nature*, *308*(5955), 181–184. https://doi.org/10.1038/308181a0

Williams, G. C. (2018). *Adaptation and natural selection: A critique of some current evolutionary thought*. Princeton University Press. https://doi.org/10.1515/9780691185507 (Original work published 1966)

Wilson, D. S. (1975). A theory of group selection. *Proceedings of the National Academy of Sciences*, *72*(1), 143–146. https://doi.org/10.1073/pnas.72.1.143

Wilson, D. S. (2015). *Does altruism exist? Culture, genes, and the welfare of others*. Yale University Press.

Wilson, D. S., & Sober, E. (2002). Reply to commentaries. *Philosophy and Phenomenological Research*, *65*(3), 711–727. https://doi.org/10.1111/j.1933-1592.2002.tb00237.x

Wilson, R. A. (2005). *Genes and the agents of life: The individual in the fragile sciences biology*. Cambridge University Press. https://doi.org/10.1017/CBO9780511807381

Wittgenstein, L. (1968). *Philosophical investigations*. Basil Blackwell. (Original work published 1953)

Wynn, K., Bloom, P., Jordan, A., Marshall, J., & Sheskin, M. (2018). Not noble savages after all: Limits to early altruism. *Current Directions in Psychological Science*, *27*(1), 3–8. https://doi.org/10.1177/0963721417734875

Wynne-Edwards, V. C. (1962). *Animal dispersion in relation to social behavior*. Oliver and Boyd.