TESIS DOCTORAL

THE ROLE OF COMMITMENT IN THE EXPLANATION OF AGENCY: FROM PRACTICAL REASONING TO COLLECTIVE ACTION

MIRANDA DEL CORRAL DE FELIPE

D.E.A. EN FILOSOFÍA

DEPARTAMENTO DE LÓGICA, HISTORIA Y FILOSOFÍA DE LA CIENCIA

FACULTAD DE FILOSOFÍA

UNED

2012

DEPARTAMENTO DE LÓGICA, HISTORIA Y FILOSOFÍA
DE LA CIENCIA

FACULTAD DE FILOSOFÍA

THE ROLE OF COMMITMENT IN THE EXPLANATION OF AGENCY: FROM PRACTICAL REASONING TO COLLECTIVE ACTION

MIRANDA DEL CORRAL DE FELIPE

D.E.A. EN FILOSOFÍA

DIRECTOR: JESÚS ZAMORA BONILLA

CODIRECTOR: FRANCISCO ÁLVAREZ ÁLVAREZ

Agradecimientos

Terminar esta Tesis ha sido, sin duda, gratificante. Escribirla, sin embargo, ha sido una tarea ardua, exasperante y solitaria. A lo largo de los años que ha durado esta carrera de fondo he recibido el apoyo incondicional de varias personas a las que me gustaría agradecer el haberme animado cuando me he encontrado desanimada, el haberme dado un empujón cuando estaba atascada, y el haberme alentado a seguir incluso en momentos en los que no he sabido explicarme qué hacía yo escribiendo una Tesis. En primer lugar, estoy agradecida a Paco Blanco porque, a su pesar, es el responsable de que encontrase una vocación que me llena y me hace feliz. Agradezco a mis compañeros de Departamento David, María y Lilian, así como a mi director de Tesis, Jesús Zamora, y a mi co-director, Paco Álvarez, que me hayan guiado, apoyado y confiado en mí en todo momento. Asimismo, agradezco a la UNED que me haya dado la oportunidad de financiar mi carrera académica mediante una beca FPI, gracias a la cual he podido asistir a muchos congresos, realizar estancias en las que he aprendido muchísimo, y dedicarme plenamente a hacer aquello que más me gusta. También quiero agradecer a Jose y Dani el haber soportado de manera estoica, incluso heroica, mis cambios de humor, los cuales han rozado la ciclotimia. Gracias, Juanlu, por apoyarme en todo momento, por estar ahí cuando yo no he estado, por discutir conmigo los puntos de la Tesis que me han tenido bloqueada, y por respetar y comprender los tiempos extraños que conlleva escribir una Tesis. Por último, quiero agradecer a mis padres el haber estado a mi lado todo el tiempo. Desde que les manifesté mi deseo de estudiar filosofía, hace 13 años, he recibido ánimo y apoyo para continuar. Y gracias a mi hermana Mónica por escucharme y apoyarme en todo momento, especialmente en las horas bajas. Os dije que algún día terminaría.

ÍNDICE

Introduction....9

Part I. Individual commitment....19

Chapter 1. Practical commitments: intention and control....27

1.1. Beliefs, desires, intentions....28
1.1.1. Reductionist accounts....28
1.1.2. Non-reductionist accounts....32
1.1.3. Arguments for rejecting reductionism....34

1.2. Intention and volitional commitment....38
1.2.1. Self-control....42
1.2.2. The strength of commitments....48
1.2.3. Weakness of will....53

Chapter 2. The normativity of practical commitments....63

2.1. Reasons, intentions and practical reasoning....65
2.1.1. Reasons for action....65
2.1.2. Judging and intending....75
2.1.3. The bootstrapping objection....81

2.2. Rationality requirements....87
2.2.1. Narrow-scope and wide-scope....89
2.2.2. Normative requirements and practical commitment....97

Part II: Individual and social commitments....115

Chapter 3. Empirical expectations....127

3.1. Why do people keep their promises?....129
3.1.1. Pro-sociality and altruistic behaviour....132
3.1.2. Sen on commitment as altruistic motivation....137
3.1.3. Socially-mediated preferences....140

3.2. Mechanisms that enable credibility and trust....143

Chapter 4. Normative Expectations....151

4.1. The source of the normativity of social commitments....151
4.1.1. Practice-based views....153
4.1.2. Expectation-based views....156
4.1.3. Reasons-based view....160

4.2. Reasons, obligations and entitlements....163
4.2.1. Propositional and action commitments....164

4.2.2. The normative structure of social commitments....172
Chapter 5: Social commitments and responsibility....185
5.1. Kinds of responsibility: six different concepts....186
5.1.1. Prospective and retrospective responsibility....187
5.1.2. Attributability and accountability....189
5.2. Criteria for attributing responsibility....194
5.2.1. The agent's capabilities....195
5.2.2. Causal responsibility....201
5.3. Responsibility, expectations and explanation....209
5.3.1. Responsibility and causal explanations....209
5.3.2. Two examples....219
5.3.3. Explaining and justifying....227
Part III: Individual, social and collective commitments....239
Chapter 6. Collective Commitments....243
6.1. Membership as a social commitment....243
6.1.1. Becoming a member....247
6.1.2. Exiting from membership....251
6.2. Practical commitments of collective agents....255
Chapter 7. Collective Responsibility....265
7.1. Attributing responsibility to collective agents....266
7.1.1. Bystanders and members....266
7.1.2. Collective responsibility: the case of collective omissions....270
7.2. Membership: the distribution of responsibility....275
7.2.1. The dilution of responsibility....275
7.2.2. Responsibility voids....280
Conclusion....285
References....291

Illustration Index

Figure 1: The enkratic requirement....116
Figure 2: The resolve and the enkratic requirements....119
Figure 3: The social commitment requirement....181
Figure 4: Membership as a social commitment....252
Figure 5: The enkratic requirement applied to a collective agent A&B....257
Figure 6: The resolve requirement applied to collective agents....259

Index of Tables

Table 1: Combinations of judgements and intentions....101
Table 2: Akrasia and three kinds of enkrasia....104
Table 3: Hamilton's matrix, adapted by West, Griffin and Gardner 131
Table 4: The Prisoner's Dilemma payoff matrix 135
Table 5: Assertoric and action commitments....166
Table 6: Safety Measures....281

INTRODUCTION

As rational agents, we have capacity to commit ourselves, to other agents, to groups. We can bind ourselves to our own future actions, not merely by modifying the environment, but by deciding to do so. We accept requests and comply with them—but we also make promises we know we will probably break. We belong to groups and institutions, and participate in collective agents. In fact, our capacity to bind ourselves, both volitionally and normatively, is vastly impressive. The aim of this dissertation is to analyse this capacity, and to argue that the normativity of commitments relies on the normativity of practical reasons.

The philosophical discussion about the role of commitments in human agency is quite recent, and it rarely occupies a central position in the debates; rather, it is used to support the relevant role that intentions, reasons, social norms, or collective agency play in the explanation of action. Metaphorically, different forms of commitment are frequently described as the "glue" that holds together different entities, such as an agent with her own future actions, an agent with another agent, or an individual agent with a collective one. This metaphor is interesting, not only for showing that a theory of action needs to include binding and relational concepts, but also because it describes quite well the state of the art in the philosophical debate about such binding mechanisms. Glue is normally used when there is a need of putting two separate objects together that, for some reason, cannot be held together without external intervention. I am interested in the reasons why different aspects of individual, social and collective agency need for that sticky entity, and in the relation between these aspects. In this dissertation, I will argue that the capacity for

commitment includes both an empirical and a normative dimensions. These two aspects have been usually kept separately, but they are closely related, as Searle points out:

> The notion of commitment is so crucial for understanding language, and indeed social ontology in general, that I need to say a little more about it. There are two closely related aspects in the notion of commitment, or one might say there are two components to the meaning of "commit", or even there are two senses of the word. In one sense to be committed to something is to have undertaken something in a way that makes it difficult or awkward to change course. For example, in this sense we speak of a general as having committed his troops on the left side of the front. There is a type of irreversibility to this sense of commitment. The other sense of "commitment" involves an obligation, or other deontic requirement. If I have made a promise to come and see you, then I have undertaken a commitment to come and see you. The way in which these two senses often coalesce is that sometimes it is a result of some action of an irreversible kind that I place myself under an obligation. And having placed myself under an obligation I have in a sense created an irreversible course. Promising, for example, contains both irreversibility and deonticity, and promising is a paradigmatic form of commitment. [1]

Committing oneself seems paradoxical: normatively, one has the power to release oneself, and thus the normative binding is suspicious; empirically, the reasons for blocking future paths of action are not easily explainable. And the same goes for committing to others: why do people acquire a voluntary obligation towards others? And why do they live up to their commitments?

Moreover, the concept of commitment has been used with many different meanings, although they all point at a set of similar phenomena: bonds to consistent courses of action[2], altruistic actions[3], strategic behaviour[4], intentions to perform an action in the future[5]. Certain kinds of commitment will not be explored in this work—for instance, commitment to a moral value or ideal, such as friendship; commitment as the acceptance of a social norm; commitment to a belief of any kind. My focus will be those commitment that stand in a particular relation with normative reasons for action.

1 Searle (2007: 16–17)

2 This use was introduced by Becker (1960), and has been mainly developed in organizational theory and social psychology.

3 Sen (1977); (1985); (2005) is the most salient proponent of this use of the concept.

4 This use was first introduced by Schelling (1960); (2007).

5 This use is widely accepted in philosophy of action; it was introduced by Bratman (1987).

Tuomela offers the following definition of commitment, which pretty much expresses the concept I will use in this work:

> [C]ommitment primarily means being bound to something in a way that gives a sufficient reason for action related to the object of commitment. [6]

This definition is not very precise, nor reveals anything about the nature of commitment, except for one thing: that there is a justificatory, or at least explanatory, relation between the fact that one is committed, and the object of the commitment. Commitment and reasons are therefore closely related. When we have reasons to do something, we tend to do that something, and having those reasons block other possible alternatives for which we have less powerful reasons. Similarly, those reasons have certain normative force: acting against our reasons, this is, being *akratic*, is a failure of rationality.

This dissertation is divided in three parts, following a distinction stated by Castelfranchi[7]. He argues that commitments are present at three levels of agency: individual, social and collective. It is a methodological, rather than ontological, distinction. The link between each of these three levels has not been studied in depth—this is the goal of the present work.

First, reasons impose certain constraints, which are both volitional and empirical, to our actions. To be committed to a goal entails to hold the belief that certain facts require, or justify, the goal on question. We take these facts as reasons to achieve our goal —for example, I am committed to finish writing this dissertation, because I find that this is what I ought to do, given the reasons I have[8]: I believe that it will improve my happiness, which is something I believe I ought to do, in general. But not only I have reasons to finish this dissertation: I also *intend* to do so, and I have the volitional capacity to control my behaviour in the light of my intentions. While I keep this intention, any action that makes my intention infeasible, such as spending the summer lying on the sand in a wonderful beach, is an instance of procrastination, or even weakness of will. Hence,

6 Tuomela (2007: 27)

7 Castelfranchi (1995).

8 I should say, though, that my reasons for writing and finishing this dissertation have varied in strength and content, specially at the final stretch of the writing process.

our intentions are normatively constrained by our normative judgements about what we ought to do, or what we have most reason to do. Our actions, on the other hand, are normatively required by our intentions. This is the basic structure of *practical commitments*: decisions to do something in the future because of some reasons we consider that make it the case that we ought to do that something.

Second, promises and requests are very similar; normally, when an agent commits herself to another agent, she also adopts a practical commitment. What kind of reasons are implied in this commitment? We can promise to do something that, had we not made that promise, we would not find any reason to do. Yet *now* we have a reason. This is so because promises, requests, agreements, and other kinds of *social commitments* create reasons for action, while practical commitments only acknowledge the reasons the agent previously had. This magic creation of normative reasons from a simple social interaction has puzzled many philosophers for many decades[9]. Several solutions have been offered, although there is still a strong disagreement on which of the options offers a better explanation of self-imposed social obligations. I will argue for a solution based on the foundations of social commitments, which I claim to be practical commitments[10]. This is: socially acquired obligations also relate to reasons, just as practical commitments; the capacity to create a new reason for action, I will argue, is an exercise of our normative powers, which include possessing things, giving a gift, owing things to others, and so on.

Third, many agents can socially commit themselves to the others to achieve a shared goal that requires to be collectively achieved. Through their commitment, individuals become *members* of a group which is itself capable of acquiring practical commitments, just as an individual agent is. However, collective agents cover a wide range of phenomena, from two persons walking together to a transnational corporation, for

9 In fact, this is Hume's problem of promising, stated in his *Treatise of Human Nature* [1739] (2007: 1:).

10 It might be well that things work the other way around: we first (evolutionarily speaking) acquired the capacity to commit ourselves to others, and later we developed an internal conception of self-imposed obligations. In any case, it does not affect the argument, because each commitment (practical and social) can be analysed in terms of the other. My point is, precisely, that we do not need to appeal to morality or social conventions in order to explain the normative structure of social obligations.

example. I do aim to provide a general definition of what a collective agent is—this task would largely exceed the scope of this work. Thus, I will limit my analysis to the explanation of membership in terms of social commitments, and suggest that, as long as the group has the necessary mechanisms for collectively acknowledging and assessing reasons, it is capable of acquire practical *collective* commitments to do something in the future.

This work is structured around these three claims, each of which is analysed in a different part. Each part is preceded by an introduction serves as the red thread that constitutes the framework of the overall argument. Those three introductions try to connect the ideas expressed in previous Chapters, and to bring forward the contents of the following Chapters. Thus, the words that follow aim to introduce the questions I will address in each Chapter.

This dissertation begins in Chapter 1 with a brief overview of the philosophical debates around the nature of intentions (§1.1). The reason why I introduced this discussion is the following. Not only it would be good, or valuable, or useful, to have the capacity to commit ourselves: we do have this capacity. And the explanation of why this is so lies in the nature of intentions, particularly, in their capacity to exert control over our behaviour and decisions. Thus, are intentions a proper and irreducible mental state on their own (§1.1.2), or are they a kind of belief, desire, or a compound of those two mental states(§1.1.1)? I will argue, against reductionist claims, that intentions have to be considered as functionally different from beliefs and desires, due to their capacity to control our present and future actions (§1.1.3). However, what is it meant by self-control, or volition? If we were not capable of controlling ourselves, we would lack intentions: we would only have certain beliefs about what we will do in the future, considering as evidence the desires we foresee we will have at that moment (§1.2). Nevertheless, rational agency is temporally extended. We have the capacity to foresee, and manipulate, our future actions, even though we know that our motivations may change when the time comes—this is the function of self-control (§1.2.1). But how does this capacity affect the commitments we may have? I will argue that it constitutes the strength of a commitment

(§1.2.2). Sometimes we stick to our commitments even when it is irrational to do so, for example, in the light of new evidence pointing out that our goal is unattainable. Other times, however, we fail to live up to our commitments: self-control requires effort, and motivation might not help to maintain our intentions. This is what weakness of will is: a failure of self-control that takes place when we hold inconsistent intentions, due to the fact that one of them is not exerting the proper control over our actions (§1.2.3).

So, we have the capacity to commit ourselves insofar we can exert control over our future behaviour and deliberation processes (for instance, by blocking them). But not only we can do this: we are required to do so if we aim to be rational agents. In Chapter 2 I examine the normative structure of practical commitments. I first analyse the normative elements involved in practical commitments (§2.1). First, we are responsive to normative reasons for action, which are facts that justify an action, this is, they serve as premises when deliberating about what we ought to do (§2.1.1). What is the relation between reasons and intentions? It is widely accepted that the conclusion of practical reasoning is an intention[11]. I will argue against this thesis: the conclusion of practical reasoning, I will suggest, is a normative judgement, which is a belief (§2.1.2). In fact, reasons and intentions hold a conflictive relation, because intentions cannot create new normative reasons for action (§2.1.3). If I do not have any normative reason to eat chicken instead of lamb, do I create a new (normative) reason to eat chicken merely by intending to do so? This is the core of Bratman's bootstrapping objection[12]. Thus, intending does not have justificatory force—normative judgements, though, do have the capacity to justify and to constrain intentions, because of the normative requirements governing rational agency (§2.2). It is widely accepted that rationality imposes certain constraints to agency (in order to be considered *rational* agency); but there is little consensus about how the logical form of these requirements is better understood. The two main positions are wide-scoped and

11 The mainstream view in philosophy of action is that the conclusion of practical reasoning is an intention Brandom (1998); Broome (2002); Stroud (2003), or an action Dancy (2004a); Tenenbaum (2007), or any of them—decisions or actions Alvarez (2010b).

12 The bootstrapping objection was introduced by Bratman (1987), and similar arguments can be found in Wallace (2001) and Raz (2005).

narrow-scoped formulations (§2.2.1). I will argue that a narrow-scope formulation is preferable, and provide a formulation of the two main rationality requirements: enkrasia and resolve (§2.2.2). These two requirements constitute the normative structure of practical commitments, relating intentions, reasons and normative judgements.

The second part of this dissertation analyses the social dimension of commitment. Following the same structure as the previous part, Chapter 3 is devoted to the explanation of why do people commit themselves. Of course, it would be socially useful, and even morally good, that people kept their promises; the thing is that they in fact do so, and this calls for an explanation which cannot be based on the normative correctness of promise-keeping. The practice of promising is strategically puzzling. Commitments are needed when there is an incentive to free-ride (§3.1). Do words change this incentive? Why are commitments credible, and why are thy fulfilled? The simple answer is that social commitments have the capacity modify the expected payoffs, and so it becomes in the self-interest of the agent to fulfil them. This capacity is based on two control mechanisms: reputation and emotions (§3.2).

Again, promise-keeping and other forms of social commitments are also subject to normative constraints: this is the topic of Chapter 4. Why ought people to keep their promises, fulfil the responsibilities they have undertaken, live up to their social commitments? I analyse three alternative answers to this question (§4.1): practice-based, expectation-based, and reason-based views. Besides moral or legal obligations, I will argue, social commitments give rise to a set of rights an obligations, that are related to the rational authority an agent has over her reasons for action, and the normative powers to transfer her authority to another agent. Thus, I will argue for a reason-based explanation of the normativity of social commitments (§4.2). However, simply declaring our own intentions also give the hearer reasons to believe that we will do what we assert, but they do not generate an obligation to do so: this is why it is important to distinguish between action and propositional commitments (§4.2.1). Action commitments entail goal adoption, the creditor's uptake, and the debtor's transfer of the authority to justify her own actions to the creditor (§4.2.2). Committing oneself to others, thus, differs form

internal practical commitments in that it is no longer possible to legitimately change one's mind. The most basic right of the creditor, then, is the right to release to debtor.

Chapter 5 is analyses the relation between social commitments and attributions of responsibility. Social commitments can be seen as the uptake of the responsibility to perform an action in the future: I will argue that the normative expectations entailed in social commitments serve as a basis for responsibility attributions. I will first present the scope of responsibility (§5.1), in which two main distinctions can be made: retrospective and prospective responsibility, on the one hand, and attributability and accountability, on the other. Then, two criteria for attributing retrospective responsibility (in the sense of attributability) will be explored (§5.2): agential capabilities and causal effectiveness. Causal explanations frequently involve not only the empirical expectations about what will happen, but also what should happen, given the agents' obligations (§5.3). Furthermore, excuses and arguments for exempting the agent show the interaction between the accepted normative judgements and individual reasons, and affect our evaluative judgements of responsibility.

The third part is devoted to collective commitments and collective responsibility. The main claim of Chapter 6 is that collective agency requires a normative structure made up by social commitments of individual agents, and practical commitments of the collective agent. I will argue that collective agents, as distinct from mere aggregations of individuals, require membership as affiliation, which is a social commitment between an individual and a group agent, of between two or more individuals, who create through their reciprocal social commitments a collective agent (§6.1.1). Are members subject to obligations *qua* members? They do, insofar ceasing to be a member is analogous to being released from a social commitment (§6.1.2). While being a member, an agent becomes socially committed to promote the group's goals, accepting those goals as normative reasons for action. Collective practical commitments to a goal are analogous to individual practical commitments (§6.2). They are also subject to rational requirements, particularly, to *enkrasia* and *resolve*. Thus, the collective agent and its members incur in a rational

obligation when the collective is committed to a goal. Finally, the decision procedure the group adopts also serves as a normative reason to accept the conclusion, as it is shown by the discursive dilemma.

The aim of Chapter 7 is to apply the framework of responsibility attributions developed in Chapter 5 to collective agents. First, it will be argued that responsibility cannot be attributed to a group of aggregated individuals (§7.1.1). This would be a case of *shared individual responsibility*; collective responsibility requires a collective *agent* to be attributed. Hence, collectives, as well as individuals, must meet certain agential requirements (§7.1.2). Second, on what basis can collective responsibility be distributed amongst the group members (§7.2)? I will present three confronting perspectives on whether collective responsibility is shared amongst the members, and if so, whether it is *diluted*: the greater the group, the lesser the responsibility each member holds (§7.2.1). I will argue that insofar membership as affiliation requires acceptance, individual members always share collective responsibility: this is why mere bystanders can be individually, but not collectively, responsible for failing to act jointly. The degrees of responsibility will depend on the agent's role in the decision processes and in the actual production of the outcome: this is why the different roles an agent can have within a group are determinant. Finally, in §7.2.2, I will bring back the discursive dilemma, and examine whether inconsistencies between the individual and the collective level can produce responsibility voids in the sense of responsibility as attributability. This would require that the collective agent is responsible, but its members are exempt.

PART I. INDIVIDUAL COMMITMENT

To close your ears to even the best arguments
once the decision has been made: sign of a strong character.
Thus an occasional will to stupidity.
F. Nietzsche, *Beyond Good and Evil*, Section 107

The first part of this dissertation is devoted to individual commitments, understood as practical and internal commitments to achieve a goal. My aim is to analyse their structure, how they are related to motivations and self-control, on the one hand, and to reasons and normative judgements, on the other. Although my main concern in this dissertation is with normativity, I believe it is important to refer to the fact that people stick to their commitments, not only because of their normative force, but also because of the conduct-controlling function of intentions.

The kind of bond that I create by committing myself to a future action is twofold. On the one hand, commitment is volitional, insofar it involves self-control and volition. My intention to obtain a PhD conditions the scope of my subsequent intentional states and action. Volitional commitment, then, expresses the control-centred aspect of intentions. It is possible to voluntarily commit our will to a goal through the performance of certain actions precisely because of our capacity to act upon previous decisions. Without this capacity, forming plans would be an outright loss of time and resources. If action is guided by out-of-control motivators (such as compulsions)[1], then our capacity to

1 Although, as Wallace (1999b) points out, the assumption that motivational states are not under our direct control seems widely spread, specially amongst the internalist view of reasons.

stick to our plans would be limited to those cases in which our plans luckily coincide with our passively suffered passions. Happily for my aim to obtain a PhD, I believe that we are planning creatures because we have the capacity of self-control, this is, the capacity to commit ourselves to paths of actions, and, through our commitments, exert rational guidance over our deliberation processes, judgements, choices and actions.

Second, committing myself to obtaining a PhD also has a normative dimension. The normativity of intentions is related to the reasons for intending, to the judgements we make about what actions we ought or ought not to do, given the reasons we have, and with the bond that these reasons ans judgements have with our actions, on the one hand, and our other intentional states, on the other. The main idea underlying the normative dimension of intentions is that our commitments are reason-responsive. We are able to justify and explain our actions, and to infer what is rationally required from our commitments.

In sum, practical commitments have two different but related dimensions: volitional and normative. This distinction was first proposed by Bratman[2]. Watson[3] also draws the distinction between volitional and normative commitments. He claims that volitional commitment comes into play when there is a need to choose (i.e. to form an intention) and we lack a judgement on what to do; normative commitment would in turn be a bond between the reasons we have for φ-ing and the judgement that we should do φ, while. Finally, from a different theoretical perspective, Mele[4] suggests a similar distinction. He proposes that practical commitments (which would be commitments to engage in a course of action) involve both evaluative and executive commitments. Evaluative commitments are commitments to our judgements about what we ought to do, what is the best possible path amongst the alternatives we (believe we) have. On the other hand, executive commitments guide the execution of our choices. I will broadly follow Bratman in what concerns volitional (or executive) commitments, which exceed Watson's account,

2 Bratman (1987).
3 Watnson (2003).
4 Mele (1995: chap. 4).

but I will narrow normative commitments to what normative requirements the agent is subject in virtue of her volitional commitments, as well as the rational constraints of judgements and intentions, such as coherence and consistency. This view is more similar to Watson's or Mele's.

Practical commitments, then, are a bond between an agent and a goal, which can be an outcome or an action. This bond is both volitional and normative, and usually entails a nested structure of further commitments to the sub-goals that will lead to the achievement of the main goal; this is why and agent is committed both to her goals and to the course of action leading to those goals.

In principle, every active goal the agent is pursuing entails a commitment to that goal. Sometimes, though, the agent is very weakly committed to her goal. Sometimes we choose an alternative while lacking enough reasons to do so, or because of weak reasons. Sometimes we do not have a strong preference for one of the alternatives. I can choose to have a chicken dish in a restaurant, and after being informed that they have run out of chicken, switch to a pork dish without deliberating again about what I would like to eat—I just do not care. In this case, it would be odd to state that I am committed to eat chicken for lunch, even if I have chosen to do it, and formed the subsequent intention. Here, I would say that I am in fact committed; but that my practical commitment is very weak indeed, because it is not supported by strong normative reasons, nor is the product of a strong motivation. Therefore, commitment would be present every time the agent has an active goal, but it comes in degrees.

The strength of commitments is derived both from their volitional and normative dimension. An agent can be strongly motivated to achieve a goal, and can also believe that she ought to achieve that goal, everything considered, and actively refrain from reconsidering her goal. Sometimes, motivation and normative judgements go hand in hand. Other times, they do not—and this divergence explains some failures of rationality, such as akrasia and weakness of will.

My aim in the following two Chapters is to analyse the structure of practical commitments; the main argument of this dissertation is that this same structure is also

present in social and collective commitments, and is able to explain their normative structure. The structure of practical commitments can be understood in terms of the relations that hold between their different elements. Once a commitment is set up, the agent is subject to certain constraints regarding her intentions and normative judgements. A practical commitment is set up when a choice is made: the agent decides what goal she will achieve—and, in most cases, how she will do so. This choice poses some restrictions to future choices, and it is restricted by the reasons the agent has for acting. These restrictions are rationality requirements: they state what the agent ought or ought not to do given her normative beliefs and her intentions, in order to keep a *rational balance*[5] amongst her mental states. There are many levels in which rationality requirements apply. For example, regarding theoretical reasoning, rationality requires that an agent does not hold contradictory beliefs. My focus here will be those requirements over practical rationality. In particular, I will argue that there are two basic normative requirements that rationality imposes over agency: *enkrasia* and *resolve*. The first of them requires coherence between reasons for action and intentions. The second concerns the consistency of intentional states. However, the formulation of these rationality requirements will be addressed in the last Section of this first part (§2.2). They represent, I will argue, the normative structure of practical commitments. But before entering into the problem of the formulation of this structural requirements, I will discuss the elements of practical commitments, which are bound through this normative structure. This first part is divided in two Chapters, which, in turn, are divided in two Sections each. Chapter I deals with the volitional aspect of practical commitment, and Chapter II is concerned with the normative elements of practical reasoning and agency.

Chapter 1 is devoted to an overview of the different approaches to the ontology of intentions, which are divided in two main groups: reductionist (§1.1.1) and non-reductionist (§1.1.2) accounts. I will argue that it is appropriate, from a methodological point of view, to treat intentions as non-reducible to beliefs and desires,

5 See Cohen and Levesque (1990); for a more recent formulation, see Broome (2001b).

due to their conduct-controlling function, which cannot be easily explained in terms of the functions of beliefs and desires (§1.1.3). This is also why I will not formulate the normative requirements of practical rationality in terms of the requirements of theoretical rationality, which concern the appropriate relation among beliefs.

In §1.2, I will analyse the volitional aspect of practical commitments, their conduct-controlling function. I will suggest a three-staged model of practical agency, and argue that the volitional aspect of practical commitments guides the transition between choices and actions, and is also able to prompt further deliberation—for instance, deliberation about what would be the best means to achieve the intended goal. Volitional commitment is the binding force of intentions (§1.2.1); without this capacity, deliberation about future intentions would be relegated to a mere exploratory exercise about what our future actions will be. Furthermore, the capacity of self-control in not only involved in the causal effectiveness of intentional states over actions, but is also able to block further reconsideration of the goals intended (§1.2.2). Finally, I will argue that weakness of will can be analysed as a failure of the controlling function of intentions (1.2.3). A weak-willed agent intends incompatible goals, and thus the monitoring and controlling functions of her intentions are not working properly.

Chapter 2 is devoted to the analysis of two normative elements of practical commitments: reasons and normative judgements. I will argue (§2.1.1) that reasons are facts, and that they are perspective-dependent. On the one hand, what justifies, motivates or explains an action, from the point of view of the agent, is not a belief, but the content of that belief, which the agent takes as true. On the other hand, a fact is not a reason by itself: it is its role in practical reasoning what confers it the status of a *reason*. Insofar practical reasoning is always perspective-dependent, this is, that it is relative to the point of view of the reasoning agent, reasons will also be perspective-dependent; the same goes for normative judgements supported by those reasons. In fact, normative judgements, and not intentions, are the conclusion of practical reasoning (§2.1.2). I will conclude this Section by examining the relation between intentions and reasons, through Bratman's bootstrapping objection (§2.1.3), and argue that the fact that someone intends to do

something can be used as a reason in a practical inference, but intending does not create normative reasons for action.

Having analysed these elements of practical commitments—intentions, self-control, reasons and normative judgements—I will analyse the formulation of the normative requirements governing their relations (§2.2). It is widely accepted that rationality imposes certain constraints to agency (in order to be considered *rational* agency); but there is little consensus about how the logical form of these requirements is better understood, and about the reasons to prefer one formulation over another. The two main positions are wide-scoped and narrow-scoped formulations (§2.2.1). While the objectivist account of reasons and oughts does not take as valid the detachment of the conclusion entailed in narrow-scoped formulations, the wide-scope approach is subject to two problems: first, its symmetrical form does not gather the rational constraints of belief change, and second, it is subject to infinite regress, or arbitrary choice. Thus, I will argue that a narrow-scope formulation is preferable. I will show that a narrow-scoped formulation of the two main rationality requirements—enkrasia and resolve—is able to account for the two corresponding rationality failures involved in akrasia, on the one hand, and weakness of will, on the other (§2.2.2). These two requirements constitute the normative structure of practical commitments, relating intentions, reasons and normative judgements.

In order to explain the wrongness or incorrectness of, for instance, holding two contradictory intentional states, it is necessary to appeal to the normative structure of practical commitments. The wrong involved in akratic or weak-willed actions is not a moral wrong (or not necessarily), but a rational wrong. This claim is quite uncontroversial in the literature. However, I find it puzzling that this claim is not applied to social, rather than individual, commitments. The wrong of breaking a promise is usually considered a *moral* wrong, rather than a rational one. My aim in this dissertation is to argue that the violation of a social commitment, such as a promise, a command or an agreement, is a violation of a normative requirement, given both that it entails an agreement upon reasons

for action, and that it confers the creditor certain authority over the debtor's justificatory capacity. In particular, the debtor acquires the right to release the debtor, which, in the case of practical reasoning, is done through a process of reconsideration of one's reasons for action. This will be the topic of the second Part of this dissertation, while the third Part will be concerned with collective commitments, which, as I will argue, combine both practical and social commitments.

CHAPTER 1. PRACTICAL COMMITMENTS: INTENTION AND CONTROL

In this Chapter, I will focus on volitional commitments as a property of intentions. Intentions exert control over other intentions, goals, choices and actions. Commitment, in this sense, would be similar to the volitional strength[1] associated with a goal. The structure of this Chapter is as follows. First, I will examine reductive and non-reductive accounts, and conclude that, at least methodologically, non-reductive accounts are preferable. Then, in §1.2, I present a model of intentional agency consisting of four stages: deliberation, judgement, choice and action. My aim is to explore the role of volitional commitments between stages, by means of self-control, or willpower. I will suggest a characterization of self-control based in a hierarchical model of agency, and discuss the strength of volitional commitments. The concluding Section is devoted to the opposite task: to explain weakness of will as a failure to fulfil our volitional commitments, this is, a failure to exercise self-control.

1 This idea was introduced by Gollwitzer (1993).

1.1. Beliefs, desires, intentions

In the last twenty years, the concept of intention has attracted a rapidly growing interest. The focus shifted from the analysis of intentional action, or acting intentionally, to intentions as a distinct mental state. Philosophical work on intentions can be divided into two main families of views: reductive and non-reductive accounts. In what follows, I will present the main claims defended by both points of view, and then I will argue for a non-reductionist concept of (future-directed) intentions.

1.1.1. Reductionist accounts

The Belief-Desire model of agency is a simple and general model of agent, and it is based on the Humean theory of human agency. It is a refinement of folk psychological intuitions about what mental states are involved in the act of choosing, and how we explain the other's actions by appealing to their beliefs and desires. Reductionist theories of intentions claim that desires and beliefs is all it is needed to account for human action. It is possible to distinguish between three kinds of reduction: either intentions are a kind of belief, a kind of desire, or, lastly, a hybrid kind compounded by both beliefs and desires.

Belief reductionism

First, some authors, such as Velleman and Setiya contend that intention is a kind of belief. Velleman identifies intentions with beliefs that oneself will do something:

> Intentions to act, I believe, are the expectations of acting that issue from reflective theoretical reasoning. These are self-fulfilling expectations of acting that are adopted by the agent from among potentially self-fulfilling alternatives because he prefers that they be fulfilled, and they represent themselves as such.[2]

2 Velleman (1989: 98).

The account provided by Setiya is very similar to Velleman's, although he focuses on why belief is necessary for intention. The kind of beliefs in which intentions consist are desire-like, having thus motivational force:

> As Anscombe pointed out, the verbal expression of one's intention to f is the assertoric utterance of the sentence 'I am going to φ', and thus the expression of belief that one is going to φ; one cannot intend to do something without having that belief.[...] [T]he attitude of intending to do something is a matter of motivating or desire-like belief. Intention represents its object as true in the same way that belief does; under the right conditions, it will constitute knowledge. But it also motivates action after the fashion of desire. [...] Intending to φ is roughly a matter of having the desire-like or motivating belief that one is going to φ.[3]

Belief reductionism has two attractive features. First, it explains the normative requirements of practical reason by appealing to the requirements of coherence and consistency amongst beliefs. Second, it gathers the epistemic role of intentions: they provide the agent a form of self-knowledge[4], i.e. knowing what she is doing when she performs what she intends to.

The problem with these accounts is that they ultimately appeal to some motivational or desire-related attitudes, such as preferences. An intention is quite different from a mere forecast. For example, I can foresee, and form the belief that, I will be unemployed by this time next year. It is actually a quite well-founded belief; but it does not motivate me to action – except, probably, to try to avoid my fate. Velleman argues that the differences between predictions and intentions lie in what he calls the "direction of guidance". Merely predictive belief does not cause the truth of what it represents, the future state of affairs. Intentions, on the other hand, cause the truth of their content. Thus, intentions share with ordinary beliefs that they represent something as true, and aim at truth; but, contrary to predictions, they in fact cause that their content becomes true. The direction of guidance is desire-like; but it does not commit us to accept that intention necessarily involves desire[5]. The problem with Velleman's proposal is that it

3 Setiya (2007a: 663–4).

4 The claim that intentions are fundamental for self-knowledge was introduced by Anscombe (1957).

5 See Langton (2004) for an argument against the possibility of forming self-fulfilling beliefs (and still be considered beliefs, rather than hopes or wishes). Holton (2009: 18) also stresses the difficulty of forming self-fulfilling beliefs prior to having reasons that justify or support those beliefs.

requires a further distinction between "normal" beliefs with "normal" direction of fit and guidance, and beliefs as intentions, with the same direction of fit but different direction of guidance. This difference has to be based on some motivational force of the belief as intention, and thus requires to acknowledge a further component of this kind of beliefs. Hence, it does not appear very different from belief-plus-desire accounts. Furthermore, dividing the category of beliefs into two classes, depending on the direction of guidance, does not appear ontologically attractive.

Desire reductionism

Second, intentions could be interpreted as desires, or complex set of desires. This account does not deny that intentions can include the relevant means-end beliefs; instead, it claims that intending does not require believing that one will do what is intended, or that doing what is intended is possible. Ridge defines intention in the following manner:

> A intends to φ if and only if (a) A has a desire to φ, (b) A does not believe that φ-ing is beyond her control, (c) A's desire to φ is a predominant one, which is just to say that there is no desire ψ, such that A does not believe ψ-ing is beyond her control, she desires to ψ as much as or more than she desires to φ, and she believes that a necessary means to her φ-ing is that she refrain from ψ-ing, (d) A has a desire not to deliberate any more about whether to φ unless new, relevant information comes to light.[6]

Thus, intentions would be a kind of desire, standing in a particular relation with the beliefs regarding the realisation of the desire. In a recent contribution, Lemaire has proposed that intentions are indeed reducible to desires and plans:

> [A]n intention to A is a set of predominant and uninhibited desires to A, where A is a plan or at least an action. It is a complex state that encompasses a set of desires that are predominant, a gate mechanism in an open state whose function is to assess the quality of the comparison of these desires given the stakes at hand and finally a plan that is the content of the predominant desires.[7]

From this perspective, intentions can be reduced to desires and plans. Plans are the content of the desire, and might be themselves irreducible. Predominant desires are either

6 Ridge (1998) quoted in Holton 2009, 18.

7 Lemaire (2012: 21).

inhibited or facilitated by a gate-keeping mechanism. The gate-mechanism can block the fulfilment of the desire, and trigger deliberation about what to do. Uninhibited desires are those that have not been blocked. Plans, on the other hand, are necessary but not sufficient for intentional states. They are necessary because plans are the content of the desire, which is the necessary component of intentions; but they are not sufficient, because, following Lemaire, one can form plans without forming an intention: for example, deliberating about what would be the best way to go to the library, without any intention to get there (for instance, because we are asked for directions). Along these lines, Mele argues that plans constitute the representational content of the intention, but they do not suffice to form an intention, because they lack motivational force[8]. However, Mele argues for a non-reductionist account of intentions; we will turn to his theory later.

Belief-Desire accounts

Third, it is possible to defend that intentions are reducible to a set of interconnected desires and beliefs. In his well-known article 'Actions, Reasons and Causes'[9], Davidson offered a reductive theory of intentions: they are taken to be descriptions of what the agent's actions in terms of her primary reason to act. A primary reason, following Davidson, is a pro-attitude towards action that have desirable features, along with the agent's belief that the performed action has those features.

Other authors require that the agent believes that she will perform the action. In this sense, they coincide with the belief-based account we presented above. However, Belief-Desire accounts insist in that there must be an accompanying desire that motivates the agent, thus triggering action. For example, following Audi, an agent intends to perform some action A when she believes that she will (probably) do A, and she desires to A more than she desires to do anything else, so her desire has guiding force[10]. Along the

8 Mele (1992); (2009).
9 Davidson (1963).
10 See Audi Audi (1973: 395); see also Audi (1993).

same line of thought, Davis[11] argues that, when intending, an agent believes that she will A because her desire to A will motivate her to do so. Davis' strategy aims to avoid the cases in which an agent expects and wants something, and nonetheless she does not intend it – for instance, I expect and want that the sun rises tomorrow. Restricting the object of the intention to the agent's own actions does not solve the problem, for an agent can have unrelated beliefs and desires over the same object.

1.1.2. Non-reductionist accounts

This family of views is also called *functionalist* accounts, because they focus on the role that intention plays in agency, rather than on its content, or intrinsic features.

Bratman's influential account of future-directed intentions[12] stresses that intentions do not merely reflect the agent's desires. The agent becomes committed to do what she intends to do. For example, if I intend to paint my house blue tomorrow, I do not merely desire, wish or hope that I paint my house blue tomorrow. Thus, as Cohen and Levesque[13] pointed out, intention is better understood as following from *choice* and *commitment*, rather than beliefs and desires –although, of course, intentions interact with beliefs and desires, and stand in a specific normative relation to them. Commitment involves, in Bratman's theory, two dimensions: volitional and reason-centred. Volitional commitment provides the motivational aspect of future-directed intentions: "Intentions are, whereas ordinary desires are not, conduct-controlling pro-attitudes. Ordinary desires, in contrast, are merely potential influencers of action"[14]. The capacity of self-control is manifested through the hierarchical influence of future-directed intentions to present-directed intentions, which at the same time control intentional actions.

11 Davis (1984).

12 Bratman has extensively developed his theory of future directed intentions; as reference points, see Bratman (1984); (1987); (2009a).

13 Cohen and Levesque (1990).

14 Bratman (1987: 16).

Intentions are not irrevocable, and they can be the object of future reconsideration, which is part of the reason-centred dimension of intentions. Bratman argues that intentions have stability or inertia. In fact, this characteristic of intentions cannot be explained if intentions were to be reduced to beliefs and desires. The stability of intentions consists in that, once the intention is formed, the process of deliberation about what to do usually stops, and tends to resist further reconsideration. Nevertheless, having new information available, a shift in preferences, or the recall of a policy the agent previously had settled, may lead to reconsideration. The point is that an agent should have good reasons to reconsider her intentions. This is why stability is related to the reason-centred dimension of commitments: failing to fulfil this normative requirement can lead to rationality failures, such as weakness of will. Having formed an intention terminates practical reasoning, following Bratman. Furthermore, intentions can also function as initiators of practical reasoning: agents reason from their previous intentions to further intentions. For instance, we take our intended ends as premises to reason about what the best means would be. In this sense, intentions are plan-like: they do not only include a goal to achieve, but also the steps to be followed in order to achieve it[15].

The reason-centred and the volitional dimensions of intentions together account for a fourth function of future-directed intentions: inter and intra-personal coordination. They generate expectations about our future behaviour, and thus allow for forming further intentions that are not incompatible with the ones we previously had. In the same way, they allow for inter-personal coordination: they generate expectations on other agents, and they allow for joint action, as I will argue through Chapters 3-6.

Mele[16] also highlights the role of intentions as source of stability, as well as starters and terminators of practical reasoning. Following Mele, the representational content of intentions are plans, which can vary from the simplest cases, such as the intention to perform a single action (for instance, "I will raise my arm as soon as I see a taxi"), to more

15 This claim is controversial, although I will not argue against it in this work. I believe it is possible to have an intention and not to know, or have not decided yet, the means to achieve it. The question would then be whether this would be an intention or a (not intended yet) goal.

16 Mele (1992: chap. 8–11); see also (2003a).

complex and nested plans, such as studying a MA in Philosophy. Future-directed intentions, then are to be understood as "executive attitudes towards plans"[17]. Based on a distinction made by Searle[18], Mele distinguishes between an attitude's representational content and its psychological orientation. Believing, desiring and intending, amongst others, are psychological orientations towards representational contents. Mele argues that the executive dimension of intentions is intrinsic to the attitude orientation of intending, which consists in being settled on executing it. It is possible, thus, to desire to execute a plan, or admire a plan, without intending to execute it. Mele's account is non-reductive because he argues that the desiring attitude does not entail settledness, while the intending attitude does. Nor this settledness can be captured in terms of desires plus a belief that one will perform what is intended. Settledness has a need for a psychological commitment to the performance of the plan. Between having a desire and forming an intention, the agent has to decide to perform the action in question.

1.1.3. Arguments for rejecting reductionism

I will now suggest two arguments for preferring non-reductionist over reductionist accounts. The first reason is that reductive accounts face some difficulties that can be overcome by acknowledging intentions as a distinct mental state, at least from a functionalist perspective. Second, I believe that the dilemma between reductive and non-reductive accounts only arises when the focus of the discussion is the ontological status of intentions, not their role or their function. Insofar I am not aiming to provide an ontological analysis of mental states, but an analysis of the commitment implicit in intentions qua mental states (simple or compound), I feel free to adopt the most methodologically advantageous position, which I believe to be a non-reductionist stance.

17 This expression is used in Mele (1992); (2009).

18 Searle (1983).

Reductive accounts are problematic for various reasons. In general, they overcomplicate either the ontological qualities of beliefs or desires, or the relation between those two states. For the belief-based reductionist, some beliefs need to hold some motivational force in order to prompt action. As we have seen, this can be done through desire-like features of beliefs, or a shift in their direction of guidance. This complicates the standard view on beliefs, as opposed to desires in what concerns their motivational strength. In addition, some authors point out that belief-based accounts reduce intentions to a kind of wishful thinking[19], for the agent believes she will do what she intends based on insufficient evidence, and gaining confidence only by forming an intention.

Desire-based reductionism, on the other hand, faces the problem that either the concept of desire at stake is too narrow, or it is far too broad[20]. As Schueler[21] points out, from the broadest perspective, it is not possible to act intentionally without any motivation supporting that intention. But narrowing the concept leads to leaving aside motivational reasons that do not respond to pleasure or any other characterization of desire as an emotion-based experience. It is widely accepted that beliefs do not have motivational force[22]. Whilst the concept of belief has attracted wide attention amongst philosophers of action, specially due to its connection with the normative requirements of practical reasoning, the concept of desire lies in more intuitive assumptions. There are two main views on the concept of desire. One conception of desire is related with its capacity to initiate action[23]. From this perspective, desiring p is to be disposed to perform any action the agent believes it leads to p. This conception of desire is related to the concept of preferences in rational choice theory. Preferences lead to action in decision contexts. However, from this perspective, desires are relegated to the role of trivial motivators of

19 This argument has been presented by Paul (2009).

20 For an overview on the concept of desire, see Schroeder (2004).

21 Schueler (1995).

22 I am not claiming that this is the case, just pointing out that the motivational force of intentions, in reductionist theories, emerges from desires, or from desire-like beliefs (which are suspiciously similar to desires, at least in a broad sense). However, it is perfectly possible to attribute motivational strength to beliefs. For instance, it has been argued that beliefs about one's capability to perform certain action can motivate the agent to perform the action in question Eccles and Wigfield (2002).

23 This view is defended by Smith (1987).

action. They are trivial in the sense that, as long as choice (or decision) is identified with desire, it is merely a triggering attitude, but the question on how and why certain desires motivate remains unanswered.

Belief-desire based reductionism partially solves some of the problems inherited from belief-based and desire-based reductionism, although it faces the problem of determining the appropriate relation between beliefs and desires in order to produce an intentional state. Critiques to this account focus on two main types of counter-example: first, cases in which an agent believes she will do φ, desires to do φ, but nonetheless does not intend to φ; second, cases in which an agent has an intention, and nonetheless she does not believe that she will do what she intends. Regarding the first type, Bratman offers the following counter-example:

> Suppose I have a fleeting craving for a chocolate bar, one which induces a fleetingly predominant desire to eat one for dessert. And suppose that just as fleetingly I notice this desire and judge (in a spirit of resignation perhaps) that it will lead me so to act. But then I stop and reflect, recall my dieting plans, and resolve to skip dessert. On the present desire-belief account I had a fleeting intention to have a chocolate dessert. But I am inclined to say I had no such intention, for I was never appropriately settled in favor of such a dessert.[24]

The cases that pose a challenge to this view are those in which an agent has a desire to φ, and a belief that she will φ, but does not form an intention to φ. It is possible to desire something, and to believe that one will act upon that desire, and nonetheless not to form an intention – for example, because one is resolved to refrain from that desire, but is aware of its strength and motivational power. However, it can be argued that reductionist accounts aim to explain intentions in terms of beliefs and desires. It is not entailed that, every time an agent has a desire and a belief (regardless of how they interact), an intention is formed.

On the other hand, there are examples, Bratman argues, in which an agent has an intention to φ, but does not believe that she will φ. There can be intention in the face of agnosticism about whether one will try when the times comes, or about whether she will

24 Bratman (1987: 20).

succeed when one tries[25]. Absentminded people can harbour serious doubts about being able to remember to do, when the time comes, what she intends now. Also, certain goals are indeed difficult to achieve. For example, if I intend to become a civil servant, I study for a competitive examination whilst believing that I will not succeed, given the odds (although I do not believe that it is not possible to succeed).

To sum up, the problem with reductionist accounts is that they do not seem able to explain the functions and features of intentional states appealing to other attitudes uniquely. For example, Castelfranchi and Paglieri[26] defend that intention is not a primitive term, and argue that beliefs and goals (instead of desires, although essentially playing the same role) are necessary ingredients. They also suggest that, in order to explain the emergence of the functions played by intentions, a further "fundamental property" seems to be still missing: this would be the role played by commitment.

The second reason for preferring non-reductive accounts is methodological. Even if intentions were reducible to beliefs and desires, it is not evident that it would be a good idea to do so. Reductive accounts do not seem to have an explanatory advantage over non-reductive accounts, except for that they also reduce our ontological commitments[27]. Were the connection between desires, beliefs and intentions clear, I would think reducing the number of entities is methodologically a good thing; however, the complexity of the theoretical connections between beliefs and desires in order to produce intentions makes it difficult to explain all the functions played by intentions only appealing to the other two concepts. I prefer to treat intentions as proper mental states, the capacity for committing the agent as a function, and explore their relation to beliefs and desires; but, again, this is a methodological choice, not an ontological one. In fact, functionalist and reductionist theories of intentions are not ontologically incompatible. It is perfectly possible to focus our research interest in what functions intentions play, specially as mediators between

25 Ibid., 38.

26 Castelfranchi and Paglieri (2007).

27 This, of course, needs not to be the case; it is possible to propose a reductive account of intentions in terms of beliefs, desires, plans and commitments, for example, treating these as primitive (and therefore increasing the number of fundamental entities).

practical reasoning and action, while believing (or not) that, ontologically speaking, beliefs and desires (or any other set of entities) are the basic mental states from which all the rest of mental attitudes emerge – intending, hoping, wishing, admiring, liking, wanting, being certain, having doubts. Examining the relations between the components of intention does not exclude analysing the functions of intentions qua proper (although compound) mental states. In fact, some of the functions of intentions are better understood in relation to other mental states. For instance, the normative requirements of practical reasoning are better understood as analogous to the normative requirements amongst beliefs; however, the normativity of intentional commitments exceeds the scope of theoretical reasoning[28], and thus the binding function of intentions constitutes, by itself, an object of study.

1.2. Intention and volitional commitment

It is common in the literature to distinguish between (at least) three stages in rational action. First, the agent gathers information and evaluates her options; then, she selects one of these options, and third, she initiates the action chosen[29]. This staged model broadly corresponds to the folk-psychological representation of the sense of agency. Pacherie offers the following narrative to describe this conception:

> *Conscious* deliberation on the basis of our *conscious* beliefs and desires yields a *conscious* decision to pursue a certain *conscious* goal, leading to the formation of a *conscious* intention or volition to realize that goal. Our *conscious* intention in turn causes our action, by *consciously* initiating and *consciously* controlling it. While acting we experience our *conscious* intention as causing our action and on that experiential basis, we are able to judge immediately after acting that we were the agent of the action. On that story, the causation of action by *conscious* mental states and the sense of agency for actions are but two sides of the same coin.[30]

28 Specifically, I have in mind the kind of normative bonds between judgements about what one should do and the formation of an intention, or between intending an end and intending the means. I will explore further these bonds in §2.2.2. I am not claiming, however, that these normative bonds are a proof of the non-reducibility of intentions, for desires could play that role (although I do not know how).

29 Explicit formulations of the stages involved in rational action can be found in Gollwitzer (1990), Kalis et al. (2008), and Holton (2006); (2009: chap. 3).

30 Pacherie (2011: 442), her italics.

Although I am not assuming that every step is necessarily conscious, this narration illustrates how agents experience the transit between each of the stages. The kind of control that intentions exert over other intentions, plans, goals and actions is perceived by the agent; it is precisely because of this sense of control that the agent knows that she is performing the action.

Practical reason is the capacity to respond to normative reasons for action[31]. It can be viewed as a mechanism to modify the motivational forces of the agent[32]. Rational agency, on the other hand, is the capacity to reason and chose according to one's preferences and beliefs[33]. It is also claimed that rational agency, in this sense, involves rational guidance: "the fact that our behavior is controlled by our deliberative understanding in cases in which we succeed in complying with our judgments about what there is reason to do"[34]. The model I want to explore in this Section is both a model of practical reason and rational agency, because the bonds between the three stages involve both a normative dimension, which will be explored in Chapter 2, and a hierarchical control from deliberation to action, which is the topic of the present Section.

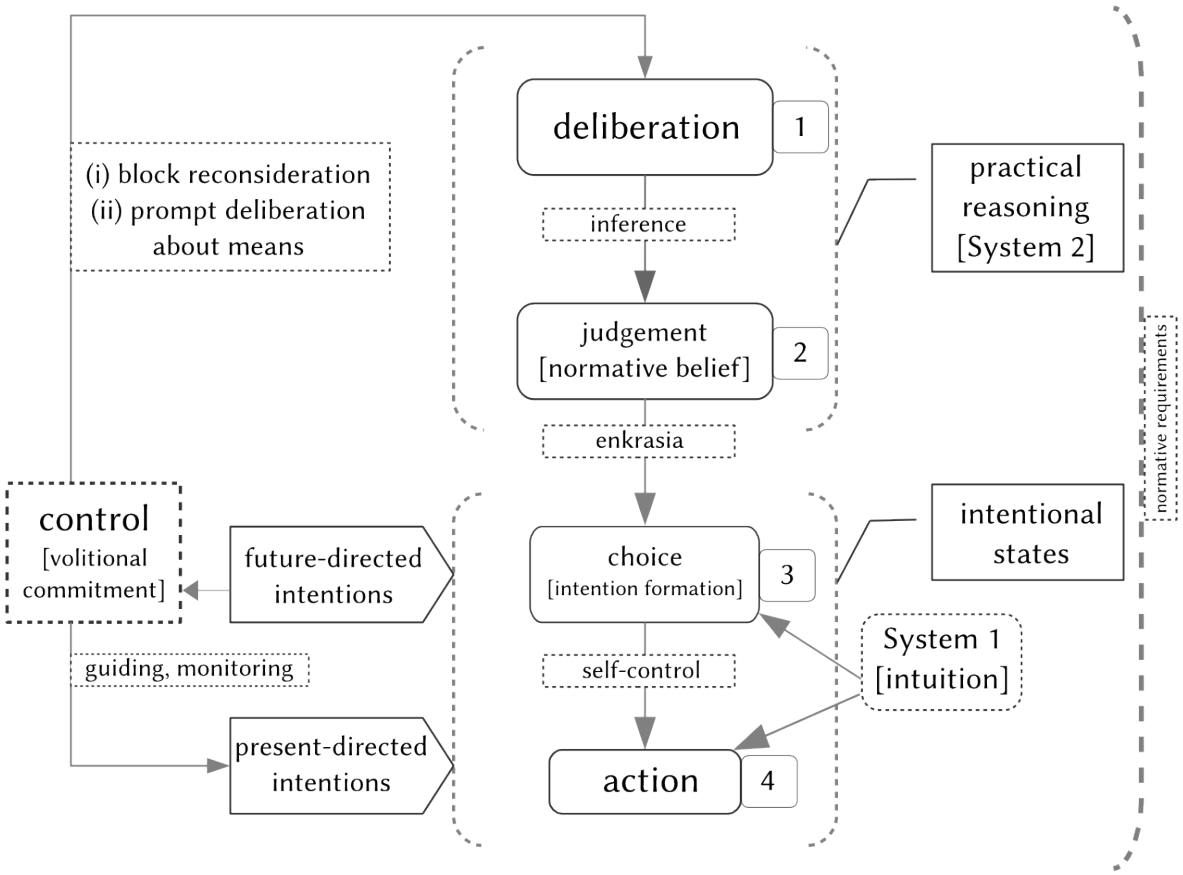

Figure 1. Three-staged model of practical commitments

The first two stages form the practical reasoning process: the enquiry about what one should, given the reasons one has. Together, they form the practical reasoning process. The step from deliberation to judgement is inferential; what inferences the agent makes depend on her deliberative commitments, a set of goals, values and norms the agent is committed to.

The step from practical reasoning to the formation of an intention is what Searle calls the first gap in rational agency: "when one is making rational decisions, there is a gap between the deliberative process and the decision itself, where the decision consists in the formation of a prior intention"[35]. This gap, Searle argues, consists in that, when making a choice, an agent always sees alternative options to choose, and they sense themselves as causally efficacious in the choice made. This is, the agent does not feel that, given the circumstances, she could have not chose otherwise: in this case, it would not be a "real" act of choice.

To make a choice (stage 3) is to make the decision to act. In this sense, judgement and choice are two distinct kinds of decisions. Judging that we should φ is equal to *deciding that* we should φ; choosing to φ is *deciding to* φ, this is, forming the intention to φ[36]. Similarly, Mele[37], following Kaufman[38], distinguishes between "practical deciding" (deciding to act) and cognitive deciding (deciding that something is the case). However, while "decisions that" something is the case (for instance, that I should work in my dissertation at least six hours a day) are the conclusion of a practical reasoning process, "decisions to" perform an action (i.e. forming the intention to work seven hours a day in my dissertation) are intentional actions themselves. Future-directed intentions are mental states that are actively formed[39], although the act of choice is not necessarily coming from

35 Searle (2001: 62).

36 Holton extensively argues for this distinction in Holton (2009: chap. 3).

37 Mele (2003a).

38 Kaufman (1966).

39 I prefer to narrow the scope of the claim to future-directed intentions because, regarding present-directed intentions, the action of forming the intention and the initiation of the action itself are temporally overlapped, and exploring that conceptual differentiation would exceed the scope of this work. Thus, I do not mean that future-directed intentions are different from present-directed intentions in this sense; I leave that question

a practical judgement. Habits[40], or expertise on certain area[41], make people choose without undertaking an explicit reasoning process leading to the formation of that intentions. Dual-process theories of thinking explain this double origin of choice by appealing to two different systems: System 1 and System 2[42]. This family of theories distinguishes between fast, heuristic, associative and implicit operations, which would correspond to intuition (System 1) and slower, conscious and deliberately controlled processes, this is, reasoning processes (System 2). The interaction between those two systems is complex, and whether the agent chooses by means of one or another system depends on habits, familiarity with the context, and the levels of uncertainty and risk under which the choice is made, amongst other factors. In fact, many cognitive biases in judgement and choice find can be explained through the tendency to think intuitively rather than following a reasoning process.

Volitional control has two directions. First, future-directed intentions have the capacity to control the formation of further intentions, as for example in choosing the appropriate means to achieve an end the agent is pursuing. In this sense, stage 3 is connected to stage 1: intentional states prompt deliberation about how to achieve their goal. Second, future-directed intentions also exert control over the agent's behaviour, monitoring present-directed intentions and actions. Thus, volitional commitment also plays an important role in the transition from stage 3 to stage 4. The aim of this Section is to explore what volitional commitment consists in. First, I will present self-control as a necessary capacity for committing oneself to future courses of action. I will argue that volitional commitment is a property of intentions that display the capacity for self-control. Then, I will discuss the graduality of intentional commitments, this is, their strength. I will present resolutions as a kind of intentions that are formed precisely to ensure the transition between stages under the control of that resolution, specially by

open.

40 Mele (2009).

41 Klein (1999).

42 See Stanovich and West (2000); Kahneman (2003); see Evans (2008) for an overview of dual-process systems.

means of blocking reconsideration. I will finally present weakness of will as a failure to maintain a volitional commitment.

1.2.1. Self-control

Intentional agency requires the capability of causing one's actions; this capacity has been analysed as volition, willpower, self-regulation or self-control[43], amongst others:

> To be an agent is to be able to intentionally make things happen by one's actions. The integrity of agency will be essentially diminished, if an agent cannot volitionally make his own decisions or choices, perform actions in accordance with his own will, or successfully carry out intentional actions to their completion. [44]

Volition, however, is far from being a unitary concept: it is related to decision-making, to acting according to one's intentions, to the initiation of the action, or to the feeling that we are causally effective over our decisions, intentions and actions[45]. Some authors emphasise the role of volition in one or more of the spheres mentioned above; others argue that the concept of volition is reducible to other more basic concepts such as "trying", which would be another name for initialising a proximal intention [46]. I believe that the concept of volition used in the philosophical literature to address the act of the will involved in intentional action cannot be understood independently of the experience of control that we have over intentional processes, the sense of causal effectiveness. Volition is a phenomenological concept, and is related to the capability of self-control:

> By "volition" here I mean a kind of motivating state that [...] are directly under the control of the agent. Familiar examples of volitional states in this sense are intentions, choices, and decisions. It is distinctive of states of these kinds that we do not think of them as belonging to the class of mere events in our psychological lives, along with sensations, moods, passing thoughts, and such ordinary states of desire [...]. Rather intentions, decisions, and choices are things we do, primitive examples of the phenomenon of agency itself. [...] The difference, I would suggest, marks a line of fundamental importance, the line between the passive and the

43 See, for a recent overview on self-control and willpower, Henden (2008) and Sekhar Sripada (2010).
44 Zhu (2004: 179).
45 Roskies (2010).
46 Adams and Mele (1992).

active in our psychological lives. [47]

It is a curious fact that the concept of control can be used in a reflexive sense. I can be in control of a car, of a situation, and even of someone else. In the car example, to be in control means to be driving it: without me exerting control over it, the car would not move. Controlling a situation entails to manipulate it to avoid deviations from my goal. To be under the control of someone means that we cannot exercise our agent capabilities, for anything we do that does not go in accordance with our controller's goals will be corrected[48]. Exerting control, thus, consists in manipulating the object for it to behave as we want it to behave[49]. In this sense, the possibility of self-control gives rise to two questions: when I exert self-control over myself, what is the object of my control? And, who is the subject? The natural answer is that I am the object of my control, and I am the subject as well. But this leads us to the second question: why would I need to control myself in the first place?

In fact, in the psychological literature, self-control refers to a conscious and effortful "capacity for altering one's own responses, especially to bring them into line with standards such as ideals, values, morals, and social expectations, and to support the pursuit of long-term goals"[50]. In general, this alteration of one's responses is done through inhibition, on the one hand, and the initiation of an alternate course of action, on the other. Self-control is not limited to "resisting temptation": it is also involved in other

47 Wallace (1999a: 636–7).

48 It is also possible that an evil genius, or a mad scientist, manipulate our desires and beliefs in order to produce the behaviour intended (by them). This is also a form of control, although much more subtle, so I cannot realise that I am being manipulated. I will let this possibility aside for now, for I believe that such manipulated agent can still experience self-control, so she can normally form intentions, and act according to them. See §5.2.1 for an analysis of the so-called "Frankfurt-style cases", which reflect this kind of external control.

49 This does not necessarily entail that control cannot be exert in a context in which we lack of alternate possibilities, this is, in which a different result could not have been obtained. Fischer and Ravizza's (2000) distinction between *regulative* and *guidance* control gathers the difference between being in control and the possibility to obtain different consequences, or to act otherwise. They define guidance control as the capacity to act based on our reason-responsive mechanisms.

50 Baumeister, Vohs, and Tice (2007).

processes, such as fixing attention, controlling thoughts and managing emotions. It is a limited resource: the capacity to exert self-control diminishes with its use[51].

Let's first address the problem of why we need to control ourselves. From an evolutionary point of view, an increasingly complex social environment and interactions encourage pro-social behaviour. A capacity to voluntarily control appetitive and aggressive urges increases the subject's fitness within the group. Heatherton and Vohs[52] argue that the capacity for individual self-control has been shaped through societal forces. Thus, from an evolutionary perspective, I need to control myself in order to avoid social exclusion. As Bratman pointed out, intentions allow for inter and intra-personal coordination. But while intra-personal coordination is indeed very advantageous, it is the need for inter-personal coordination what has evolutionarily shaped our capacity for self-control. Furthermore, the attribution of the capacity of self-control is central to many aspects of our social life, such as attributions of responsibility (see Chapter 5).

The other question I raised a few lines above was who is controlling who, in the context of self-control; many theories of self-control defend that it entails a manifestation of the self, guiding one's behaviour according to the values one endorses and is committed to defend, providing reasons supporting them[53]. As I will illustrate below through the example of resolutions, it is common to have a motivation towards a goal and, at the same time, to be aware of the instability of that motivation. We can foresee temptation, laziness, or overflowing activities that prevent us from achieving our initial goal. However, when this happens, we do not automatically drop that goal, or change it for another that fits better our actual preferences and motivations. As Frankfurt[54] puts it, second-order desires guide first-order desires: they control and guide the formation of intentions. A second-order desire is a desire to have a desire. In Wallace's terminology, the lack of this

51 This theory labels the process of exhaustion of willpower *ego depletion* its initial formulation can be found in Baumeister et al. (1998); for some recent research that point out some problems of this view, see Job et al. (2010).

52 (1998); see also Heatherton (2011).

53 Frankfurt (1971); Watson (1975); Schechtman (2004); Ekstrom (2005).

54 Frankfurt (1971).

second-order kind of motivations produce an "hydraulic conception of desire", in which agents are passively moved by motivations that escape their control. Wallace proposes a differentiation between two kinds of motivation: passive conditions, or desires, on the one hand, and other motivations that are under the agent's control, such as choices and decisions[55]. Along the same lines, Mele[56] argued that the motivational force of an agent's desires is not always in line with her evaluation or assessment of the content of those desires; however, judgements about what would be best to do are typically formed on the basis of the agent's evaluations and assessments. Similarly, Searle[57] argues that the agent's commitments are desire-independent reasons for action: the agent acknowledges their normative strength, but she is not necessarily motivated by them.

The idea of a hierarchical model is attractive, because its ability to account for the resistance to reconsideration of goals even when we fail to perform them, and this failure is completely due to ourselves[58]. Bringing back the problem of self-control, I (my higher-order intentions) would be controlling myself (my present motivational forces) to guarantee some degree of success in achieving those higher order intentions. Similarly to self-regulation in social interaction, I can foresee that, as the title of a paper by E. Harman nicely says, "I'll be glad I did it"[59]. In this case, rather than evaluating the social expectations on my behaviour, I try to evaluate what I expect from myself, and to fulfil

55 Wallace (1999a).

56 Mele (1987).

57 Searle (2001).

58 In spite of their initial attractiveness, hierarchical models are problematic for two reasons, as Ekstrom (2005) argues. First, agents can reflect not only about their first-order desires, but also about their second-order desires, forming third-order desires. This leads to the possibility of infinite regress: "There is no theoretical limit to the length of the series of desires of higher and higher orders; nothing except common sense and, perhaps, a saving fatigue prevents an individual from obsessively refusing to identify himself with any of his desires until he forms a desire of the next higher order" Frankfurt (1971: 16). Thus, there are no theoretical constraints to the level of reflection (and meta-reflection) that an agent can develop until she identifies herself with those desires. This links with the second objection: the level from which an agent evaluates her lower-order desires is arbitrary. If this is so, then the self, the set of values and desires the agent identifies herself with, is also arbitrary. I will not focus on these two problems here because I am concerned with the controlling force of intentions, rather than self-control as an agential capacity. My claim is that intentions are one of the possible ways to exert such self-control. More specifically, that the volitional commitment entailed in intentional states requires (i) self-control and (ii) a hierarchical structure of intentions, goals and actions. I will assume, then, that arbitrariness is compatible with that claim.

59 Harman (2009).

those expectations. In the same way, I can examine what other goals and intentions would undermine the chosen goal, and try to not to form these intentions.

I am not suggesting that self-control is only exerted when there is a conflict between motivational states; in this case, the concept of strength of will (as opposed to weakness of will) gathers the effort to choose according to one's previous resolutions, when facing a conflict of motivation, such as a temptation[60]. Many people acquires a habit to do something that was terribly costly at the beginning, such as going on a diet; as times goes by, habits usually enter into conflict with our present motivations to a lesser extent. Hence, self-control is also displayed in cases in which the agent feels no desires or motivations to do otherwise[61]. In fact, self-control is displayed in every intentional state; volitional commitment is the exertion of one's reflective self-control over one's deliberation, choices, intentions and actions. Without the capacity of self-control, planning, moving from the intended ends to the required means, forming future-directed intentions, would be worthless: even if we now preferred doing φ later, our doing φ later would not depend on our prior intention, but on our motivations at that moment. Our intentions would look very much like predictions: rather than intending to take my dog for a walk at 6 pm, I foresee that I will walk my dog at 6 pm; but this prediction exerts no control over my action of walking my dog. Thus, why bother making plans? Deliberation about future intentions would be relegated to a mere exploratory exercise about what our future actions will be. This is not the case, at least from an experiential point of view: we feel that our intentions exert some control over us.

As Bratman argued, future intentions control our behaviour in at least two different ways. First, when forming the intention to do something in the future, we

60 Holton (2003).

61 Against this claim, see Mele (1987); (1995). He defines self-control as "the ability to master motivation that is contrary to one's better judgment – the ability to prevent such motivation from resulting in behavior that is contrary to one's decisive better judgment" Mele (1987: 54). I agree with Mele in that self-control is usually made explicit in cases in which it entails a clear effort to manipulate one's own motivations. However, if self-control is only used when there are competing motivations, we would need a different concept of control to explain how we guide our deliberation processes, how we make choices, and how we control the execution of the intended action; I do not believe that the kind of control involved in those cases is essentially different from self-control as strength of the will.

commit ourselves to a plan: the steps we will follow in order to achieve our goal. Ideally speaking (from the point of view or practical rationality), people stick to their plans, unless they revise and reject them. Second, future-directed intentions monitor present-directed intentions. Following Pacherie[62], a central function of future-directed intentions (what she calls D-intentions) is to ensure the rational control of what the agent is doing. Rational control takes two forms. On the one hand, "tracking control" has to do with plan implementation, checking whether the steps of the plan are successfully carried out before moving to the next step. It also has the function of revising the plan when the goal becomes impossible through the intended plan. While being active, intentions prompt the formation of implementation intentions, which are also future-directed[63]. In this sense, intentions are self-regulatory: higher order intentions regulate and control the formation of dependent future-directed intentions. On the other hand, "collateral control" monitors the possible side effects of the realization of the plan. Those side effects can flout out the original reasons the agent had to perform the action in the first place, or contradict other intentions the agent had, specially her general policies or resolutions.

To sum up, volitional commitment is the binding force of intentions: it is an exercise of self-control over our deliberation, choices, and actions. It does so by making the decision to perform the action in question resistant to further deliberation: the agent, so to say, moves on and does not endlessly reconsider her intentions. Also, volitional commitment consists in prompting the implementation of a plan that the agent believes will lead to the achievement of her goal: deliberating about means, forming implementation intentions, and controlling the formation of adequate present-directed intentions.

62 (2008)

63 Gollwitzer and Sheeran (2006)

1.2.2. The strength of commitments

The strength of practical commitments can vary. This variation is analogous to the degree of certainty an agent can have regarding a belief. The extent to which an agent is committed to her goals depends on many factors, such as the motivational strength of the reasons she has for performing that action, their justificatory force, or the importance the goal has for the agent. Stronger commitments usually entail more degree of control, and less tendency to revise one's intention. There is a special kind of intention, suggested by Holton[64], whose role is precisely to keep the agent committed to her goal, when facing a temptation or other kind of conflicting motivation is foreseen. A resolution is "an intention that is explicitly designed to resist the inclinations that we predict we shall later feel" [65]. It is a complex intention; it comprises an intention to φ in a future time, and a second-order intention to resist deflection, this is, to resist reconsideration. Resolutions, Holton argues, are "contrary inclination defeating intentions: intentions formed by the agent with the very role of defeating any contrary inclinations that might emerge"[66]. Thus, the agent holds two goals simultaneously: the goal to perform the intended action, and the goal to persist in her intention. Resolutions are similar to what Bratman calls policies[67], which are a type of intentions expressing a general commitment of the agent. In Bratman's account, intentions, plans and policies are the three main types of intentional pro-attitudes. For instance, I have the future-directed intention to write the next Section of this Chapter before the end of this week; I also have a plan to do so, including all the necessary steps and their related intentions; and I can also have a general policy of writing everyday for five hours. Higher-order policies are the result of the agent's reflection over her motivations; in this sense, I can have a policy of trying to avoid procrastination while writing. Bratman argues that these higher-order policies are self-governing policies, because they are means by which the agent governs, evaluates and guides her motivations:

64 Holton (2004); (2009); May and Holton (2011).
65 Holton (2004: 507).
66 Holton (2009: 119).
67 Bratman (1987); (2000).

> [T]he agent's reflective endorsement or rejection of a desire can be to a significant extent constituted by ways in which her self-governing policies are committed to treating that desire over time. She endorses or rejects a desire, roughly, when relevant self-governing policies endorse or reject relevant functioning of the desire. [68]

Thus, both resolutions and self-governing policies are supposed to exert control over one's motivations. Weaker commitments can be conditioned by our future motivations: I can decide now to watch later an episode of *The Wire*, as long as I keep wanting to do so[69]. Acting upon resolutions or self-governing policies would be a case of flawless rational agency, which has been given different names in the literature: full-blown agency [70], wholeheartedness [71], agency *par excellence* [72], or autonomous agency [73].

Nonetheless, some authors, specially those belonging to the Rational Choice Theory, argue that intentions alone (even if they are very strong) do not have the capacity to affect future preferences. Along these lines, Elster [74] has argued that, in order to influence her future desires and motivations, an agent has to do something in order to modify her future incentives to perform the action. Elster introduces the concept of precommitment to express this inter-temporal bond:

> When precommitting himself, a person acts at one point in time in order to ensure that at some later time he will perform an act that he could but would not have performed without that prior act. As I define it, precommitment requires an observable *action*, not merely a mental resolution. […] precommitment may occur either by *deleting elements* in the set of feasible actions or by *affecting the consequences* of choosing them. [75]

---

68 Bratman (2000: 48)

69 It can be argued that I am not really making a decision to watch *The Wire* later on, but just thinking that it would be a good idea (although I am not committing myself to watching it). This is so because the reason to intend to watch *The Wire* is that I desire to watch it; I do not have any other reason to do so –except, probably, the incentive of acquiring the capacity to talk about that TV series and not be left aside in thematic conversations. Therefore, when the time comes, if that desire does no longer exist, or I changed my mind and prefer to read a book, it would be irrational not to revise my intention. However, I do not see why it should not count as a decision to watch The *Wire,* even if it does not strongly commit me. After all, if I failed to watch *The Wire*, being my motivation unaltered, it could be argued that I am being weak-willed, because I would be failing to act upon my intentions.

70 Bratman (2001).

71 Frankfurt (1988: chap. 12).

72 Velleman (1992).

73 Velleman (2000).

74 Elster (1979); (2000); (2003).

75 Elster (2003: 1754); his italics.

Thus, an agent precommit herself through precommitment technologies, which would be strategies to modify the future set of options, therefore manipulating her own future preferences. Elster argues that intentions, by themselves, do not have this power. Holton, contrary to that claim, defends that, in most cases, intentions are sufficient to affect one's future behaviour [76]. Of course, this is not to deny the role of precommitment technologies as strategies to control our future choice; to use a very typical example, smokers who are trying to quit actively modify her future temptations through certain strategies, such as not having cigarettes nor lighters at home. But, for example, mental strategies such as visualising a smoker's lung to cause disgust towards smoking would not count, following Elster, as precommitment technologies. Thus, following Elster, intentions would not entail volitional commitment, for they cannot exert control over the agent's future actions[77]. I believe this is not the case; intentions, as I have argued above, stand in a hierarchical relation with other intentions and goals, and they have the capability of restraining future choices. In fact, many people is capable of resisting temptation, or conflicting desires, driven by earlier resolution to do so. As van Willigenburg[78] claims, intentions exert a kind of control over the agent that is more related to the *authority* this agent has over her own actions, and from which autonomy is derived, than to control as a *way of forcing oneself* (which would be similar to Elster's precommitment):

> [Commitment] is a way of binding oneself, that is profoundly different from the self-enforcement that is sometimes introduced to countervail problems of collective action in rational choice. One could try to exert control over a person's future behaviour (including oneself) by manipulating expected pay-offs. One could organize self-punishment or other

76 Holton (2009: 10)

77 A possibility that I will not explore in this work is that the agent may be able to foresee that she will feel regret for not having done what she intended, and that avoiding this feeling can motivate her. Intentions would in this case have an intrinsic motivational force, even in the light of preference shift. Bratman has proposed that the rationality of not revising an intention partially depends on what he calls the *non-regret condition*: (a) If you stick with your prior intention, you will be glad you did; and (b) If you do not stick with your prior intention, you will wish you had Bratman (1998: 70). Although Bratman focuses on why this condition is relevant for the stability of plans and the rationality of reconsidering an intention, foreseeing feelings of regret (or, if the goal is achieved, satisfaction), can motivate the agent to stick to her plans. Thus, although I believe that intentions are conduct-controlling even in the absence of foreseen regret, this mechanism allows for intentions to volitionally commit the agent, even in the light of Elster's framework of commitment.

78 van Willigenburg (2003).

> ways of sanctioning non-cooperative behaviour. But this is not the kind of agential control over one's behaviour involved in autonomy. It is a way of *forcing* oneself instead of *obligating* oneself. Putting oneself in charge by taking a reason as conclusive, bowing one's mind to its reasonable force, is a way of obliging oneself. *Power does not obligate, authority does*. [79]

On the other hand, however strong our commitments are, they cannot exert such degree of control over our future actions, because our capacities to reconsider or drop our commitment, and our agential authority when executing the action in question, would be undermined. To put it differently: if deciding now to act later fully causes our future behaviour, what is the role of our future selves in the production of that behaviour?

> Our autonomy over future actions requires, on the one hand, that we have the power of making future-directed decisions that are effective, so that we can determine today what will get done by us tomorrow. On the other hand, our future-directed decisions must not simply cause future movements of our bodies. If they did, our later selves would lack autonomy of their own, since they would find their limbs being moved by the decision of earlier selves, as if through remote volitional control. We must exercise agential control over our own future behavior, but in a way that doesn't impair our own future agential control. [80]

Thus, even the strongest commitment has to leave room for reconsideration, as well as allowing the agent's control at the time of our action. Otherwise, we would be slaves of our own past intentions. It seems plausible, then, that the strength of a commitment is less a matter of the causal efficacy that this commitment exert over the agent than the level of resistance to further reconsideration, which is also a form of self-control, directed towards the intention itself. In fact, the control exerted by intentional commitments is often subtle. For instance, imagine that I intend to make dinner for some friends, who are coming to my house this evening. As long as I hold that intention, it controls the formation of further future-directed intentions, and also the present-directed intentions while starting to cook. If I intend to make dinner tonight, I need to check whether there is a missing ingredient, and in case it is, going to the supermarket to buy it. Also, were I invited to go to a friend's house, I would notice the conflict between making dinner at home and going somewhere else, and (in case I stick to my previous intention) decide not to go. This can be done even in the absence of being actively thinking that I intend to

79 Ibid., 134; his italics.

80 Velleman (1997: 45–46).

make dinner; this is why the kind of control my intention exerts over me is subtle, which does not mean weak, or non-existent. Resistance to reconsideration can also be an automatic process: were I invited to a party somewhere else, I can just reject the invitation in virtue of my previous plan, without entering into reconsidering my previous intention to make dinner. However, conflicting motivations make explicit the strength of the commitment, in the following sense. Cases of loss of control without reconsideration are not very common. I can intend to wake up at 6 in the morning and set up an alarm. When the alarm rings, I just stare the wall and try to wake up, but fail and fall asleep. I may even not reconsidered my intention to wake up at 6: it is just the case that my body does not respond to my will. However, a more usual scenario involves reconsidering whether to wake up at 6 was really important ("sleeping for another hour is not going to affect my schedule in a very serious way"), or was actually a good idea ("I will be so tired that I will not be productive, defeating the purpose of waking up at 6"). Then, I change the alarm clock to ring at 7, and get asleep again. Here, I have reconsidered my intention, but I have no problems of self-control, understood as reason-responsiveness (I really have reasons to stay in bed), as being able to form present-directed intentions, or as forming further intentions that would defeat. But, if I opened my eyes at 6 and, in spite of being tired, and comfy on my bed, I struggled against my desires to reconsider my intention and just wake up, I would be exhibiting a greater level of commitment to my intention to wake up at 6 than if I reconsidered it.

Reconsideration of intentions (or, more precisely, the capacity to block it) is fundamental for the commitment to be effective, this is, that the agent actually carries out what she intends. However, reconsideration also raises normative problems, such as under which conditions it is rational to revise an intention, or whether it is rational at all not to revise it. Next Section is devoted to explain in what sense weakness of will is a failure of the agent's volitional control.

1.2.3. Weakness of will

In the literature, both weakness of will and akrasia are usually analysed as the same kind of phenomena. They both refer to cases in which there is a conflict between the requirements of practical reason and action, an inconsistency amongst judgements, intentions and actions. Akrasia, as opposed to *enkrasia*, refers to the inadequate relation between a normative belief and the corresponding intention. In order to not to act akratically, an agent is required not to believe that she ought not to φ at the time she intends to φ. From the point of view of volition, the possibility of akrasia seems to suggest that it is possible to have conflicting motivations and evaluations. This links with the hierarchical view of motivations presented in §1.2.1. It is possible for an agent to judge that she ought to do something, or that she ought to feel motivated to do something, and nonetheless not to feel motivated at the level of the first-order desires. I will explore akrasia in §2.2.2, because I take it to be a rational failure that does not involve a loss of self-control, but a failure to comply with a normative requirement of rationality. Weakness of will, on the other hand, is the failure to act upon one's practical commitments, and therefore constitutes a failure of volitional commitment. It violates the requirement of consistency amongst intentional states, because when an agent intentionally refrains from doing something she still intends, she has two inconsistent intentional states simultaneously.

A preliminary remark ought to be made. I have chosen to label those two failures of rational agency following the classic names 'akrasia' and 'weakness of will'. My aim is to show that the consistency requirements between normative beliefs (about what one ought to do) and intentions are different from the requirements of consistency amongst intentional states. Traditionally, akrasia has been related to a failure to act in accordance to one's judgement (but not necessarily against one's intentions); more recently, some authors such as Holton[81] and Dodd[82] have suggested that weakness of will entails a failure

81 Holton (1999); (2004); (2009); May and Holton (2011).

82 Dodd (2009).

to act in accordance to one's intentions[83]. I believe that this distinction is relevant, and thus I have maintained the names; I could have labelled them “Failure 1” and “Failure 2”. My aim is not to show what akrasia, or weakness of will, *really* are[84], but to point out two phenomena that are problematic both for rationality and self-control.

While akrasia is the failure to hold consistent judgements and intentions, weakness of will refer to the failure to hold consistent intentional states. It is a rational failure that has its basis on a failure of self-control: the agent's intentions do not properly exert control over other intentional states. When holding an intention, certain response from the agent is required, as for example to form the relevant implementation intentions[85]. An agent displays weakness of will, Pettit[86] argues, when she holds by intentional states in the light of which a certain response (an action) is required, and nonetheless she fails to act in the required manner.

Following this account of weakness of will, consisting in holding inconsistent intentional states, it can be argued that compulsion and addiction are also cases of weakness of will. In fact, Watson[87] argues that the distinction between compulsive and weak action is normative: weakness of will entails that the agent has been overcome by desires that she ought to have been able to resist. On the contrary, in the case of compulsion, the agent could not have resisted those desires, and therefore it is not normatively expected from her to resist. However, it is not clear that compulsive actions, or those driven by addiction, amount to standard intentional actions. Imagine an agent who, despite aiming to stop consuming heroin, is unable to resist her withdrawal

83 For arguments in favour of distinguishing akrasia from weakness of will, based on the possibility of being akratic but not weak-willed, and vice versa, see McIntyre (2006) and Levy (2011).

84 Also, I believe this task is worthless, but this is a different matter. However, some authors seem to have been looking for the ordinary meaning of weakness of will, as a basis for preferring their accounts see Holton (2009); Mele (2010) criticises Holton’s account in the light of the experimental results he obtained through some surveys; May and Holton (2011) respond to Mele’s critics. I will not enter here into the discussion about what concept gathers the ordinary use of weakness of will more appropriately, so I am open to changing the names of the rationality failures I am discussing, if that makes things clearer.

85 Gollwitzer (1999).

86 Pettit (2003a). It should be noted that Pettit uses the concept of 'akrasia' to refer to this failure; I believe that the kind of irrational action he refers to is better characterized as weakness of will.

87 Watson (1977).

symptoms and ends up by taking the drug. If she lacks of self-control, then there is no self-control to exert. Assuming that self-control can be lost, or depleted, I am not sure about the extent to which it can be said that the action performed while out of control is indeed intentional. It is not my aim here to draw a sharp line between addiction, compulsion and weakness of will. The idea that addiction and compulsion undermine some of the agent's volitional capabilities, either totally or partially, is widespread[88], specially in the literature about responsibility[89]. I will assume that this is so, and that, in principle, compulsive agents, or those moved by addiction, incur in weakness of will whenever they hold conflicting intentions; but I am not claiming neither that compulsive actions are intentional, nor that they are not.

Thus, weakness of the will represent a failure of volitional (and rational) aspects of commitment because it entails that the agent intends incompatible goals, and thus the monitoring and controlling functions are not working properly. One of the goals is necessarily doomed to failure, and this is so precisely because of the agent's intentional actions: it would be a case of boycott against oneself. I will now discuss two instances of weakness of will. The first of them is what Holton strictly considers weakness of will: the failure to keep one's resolutions. The second has been more deeply analysed from the perspective of rational failures: a situation in which an agent intends an end, knows the means, and nonetheless does not intend the means.

Holton[90] has suggested that the difference between akrasia and weakness of will lies in that, while the former consists in acting against one's better judgement, the latter entails acting against one's resolutions. As explained above, resolutions are similar to Bratman's policies. They are intentions specifically formed in order to resist to further contrary intentions the agent may have: they are *contrary inclination defeating*. Thus, the effectiveness of resolutions does not only rely on their capacity to control or guide conflicting motivations; they also serve to resist further reconsideration. Following

88 See Watson (2004c) for an overview of this claim.
89 Gideon Yaffe (2001).
90 Holton (1999); (2009).

Holton, weakness of will can be understood as a failure to abstain from reconsidering the reasons we had to do something, bringing into play some reasons that the agent already judged as bad or weak reasons. In this sense, resolutions serve precisely to block further deliberation. This function of resolutions makes explicit the relation between deliberation, choice, and the maintenance of the intention. Motivation can, and does, change through time. Large plans require intra-personal coordination. If we constantly reconsidered our goals, we would take the risk of being stuck on a loop: from intention to deliberation and judgement, and back to choice and intention.

It should be noted that abandoning a resolution does not necessarily entail that the agent is being weak-willed. For instance, an agent may have a good reason to drop a resolution. Imagine that Sarah plans to go jogging the day after. She knows that it is likely that, when the time comes, she will judge that she prefers to stay at home; therefore, her intention is a resolution to go jogging. The day after, she decides not to go jogging. To assess whether she is weak-willed, it is necessary to evaluate her reasons for changing her mind. For example, it could be raining cats and dogs; in this case, she is justified in abandoning her intention of jogging. However, if she has no reason for revising her intention, but nonetheless decides not to go jogging, she is being weak-willed. The difference lies in that, in the former case, it is reasonable to revise her belief: had she had known that it was going to rain, she would not have formed her intention. It is not reasonable to blindly stick to previous commitments, by not being sensitive to relevant new facts, or to facts the agent was unaware of. In the latter situation, however, nothing makes her change of mind reasonable. This is why, Holton argues, weakness of will is irreducibly normative: to qualify an agent as weak-willed, normative questions need to be considered, concerning the agent's reasons for acting. Holton acknowledges that assessing the reasonableness of revising one's resolutions stands in very vague grounds[91]. He proposes to use a set of rules of thumb as heuristic tools for this assessment[92]. Bratman

91 Holton (2009: 75).

92 Ibid., 160.

suggested a similar proposal, the *non-regret condition* (see note 61). The idea is that it is rational to keep one's resolutions when the absence of regret is foreseen.

To sum up, Holton restricts weakness of will to cases in which the agent revises her resolutions, when she should not have done so. This is why, he argues, weakness of will is an intrinsically normative notion. In fact, Holton claims that weakness of will differs from *caprice* in that the latter entails revising one's intentions:

> The distinction between simple intentions and resolutions provides us with what we need to distinguish weakness of will from caprice. If someone over-readily revises a resolution, that is weakness of will; if they over-readily revise a simple intention, that is caprice. Consider again the vacillating diner. Suppose he has become concerned about his tendency to keep changing his mind, and so resolves to go to a particular restaurant even if another seems more attractive later on. In other words, suppose he forms a resolution to go to that restaurant. Then if he revises his intention once again, he would not merely be capricious; he would display weakness of will. [93]

Although I believe that Holton's approach is correct in what concerns the irrationality if reconsidering one's resolutions just because we are facing the temptation that the resolution is trying to resist, I disagree with him in the reason why it is irrational. Revising (and eventually abandoning) one's resolutions is a failure of rationality insofar the agent holds conflicting intentions. The rationality of changing one's mind depends on the reasons of the agent for doing so, not on whether she is revising a resolution or a simple intention. Assuming that the agent's reasoning is normal, abandoning a resolution does not entail weakness of will. However, acting intentionally, or forming an intention to act, against one's prior intentions, whatever kind they are (simple, resolutions, policies) constitutes a case of weakness of will[94].

Weakness of will is a rational failure because it violates a resolve requirement; I will analyse this aspect in §2.2.2. Now, I will argue that it is a volitional failure insofar holding conflicting intentions necessarily entails that at least one of them will not be able to exert behavioural control. Suppose that an agent intends (A) to go to the supermarket and also intends (B) not to go to the supermarket. Were intention (A) exerting control over the

93 Ibid., 77.

94 Dodd (2009).

agent's present-directed intentions, and over her deliberation about the means required, then intention (B) would be defeated, and vice versa. If the agent has conflicting volitional commitments, she is fated to violate one of them. Of course, it is possible to have conflicting motivations; intentions, on the other hand, require consistency amongst them in order to be effective (conduct-controlling) and stable. Despite the consistency requirement, it is frequent to do something intentionally, while knowing that we have a contrary intention; it is not even necessary that we revise it. Skipping diet once does not necessarily entail to have revised and abandoned the intention to go on a diet; it is an intentional action that enter into conflict with a prior intention.

The reasons why weakness of will happens are related the capacity of self-control. There seem to be three factors involved. First, self-control, or willpower, comes in degrees. Under some circumstances, people is more able to resist temptation, or to perform better at tasks involving self-regulation. This is, willpower would be like a resource of the agent; and, as a resource, it can be depleted. Baumeister et al.[95] suggest that certain tasks cause ego depletion: which would make agents less self-controlled, and therefore more vulnerable to temptations, or to intentionally act against one's previous intentions. When faced with choice situations, agents whose ego is depleted tend to choose according to their present motivation, rather than exercising self-control in order to keep one's intentions. This does not entail, of course, that they revise or abandon a prior intention: they just act intentionally against it. Second, the instability of evaluation and motivational mechanisms plays a fundamental role. Willpower, which is required for intentions to be able to exert any kind of control, is synchronic: it is extended through time[96]. It affects the formation of further intentions both through motivational and evaluation mechanisms. Preference reversals[97] are temporal inconsistencies amongst preferences, due to the proximity (either temporal or physical) of temptation, and how the agent's evaluational

95 Baumeister et al. (1998); Baumeister, Vohs, and Tice (2007).

96 See Sekhar Sripada (2010).

97 Elster (2006).

mechanisms respond to different configuration of the choice situation. Ainslie[98] claims that hyperbolic discount provides a useful explanation of why weakness of will occurs. For example, Sarah decides on Monday to stay at home the following weekend in order to work on her dissertation. As time passes and the weekend arrives, she might devalue the reward of working during the weekend and, on Saturday, she might finally decide to go to the cinema with some friends. Would this be a case of weakness of will? Given that she intended to stay at home during the weekend because it served a prior intention of her (finishing her dissertation in three weeks, for example), intentionally going out on Saturday entails holding conflicting intentional states. Choice over time, thus, is subject to certain cognitive biases and tendencies that affect our evaluation and motivational mechanisms. Holton labels this tendency *judgement shift*: "where the options are judged as close, judgements are revised to bring them into accord with desires, rather than desires being revised to bring them into accord with judgements"[99]. It is characteristic of temptation that, after succumbing, the agent comes back to her previous judgement; this is so because she has not abandoned her previous intention, the one that enters into conflict with intentionally performing the tempting action. Lastly, the third mechanism that could be involved in weakness of will is related to the dual process model of reasoning that I mentioned at the beginning of this Section. Levy [100] suggests that ego depletion may cause that agents switch from System 2 to System 1. Choosing by means of System 2 requires more effort, more attention and, in sum, its use depletes the agent's willpower. Thus, it is plausible that weakness of will, as a temporary conflict of intentional states, could be understood either as a depletion of the System 2 resources, or as a mechanisms in order to conserve these resources before depletion. The agent still has her previous intentions, but she does no longer choose according to them, but according to System 1 instead, which displays a higher tendency to satisfy one's immediate desires (broadly understood).

98 Ainslie (2001).
99 Holton (2009: 110).
100 Levy (2011).

Finally, a failure to intend the means when one intends an end can also be a form of weakness of will, insofar there are two intentional states – what the agent intends as an end and what the agent intentionally doing, instead of doing the mean – that are in conflict. However, the conflict is more subtle in these cases. A typical example of intending an end, knowing the means, and nonetheless not intending the means is procrastination[101]. Future-directed intentions are not necessarily violated when they are not performed; they are simply in stand-by, pending to be fulfilled. A procrastinator does know how to fulfil them, but does not initiate action; rather, she intentionally does something else. While it is clear that procrastination is a failure of self-control, it does not always entail weakness of will, because the task in which a procrastinator is spending her time does not always enter into conflict with her previous goal. Suppose that Joe intends to write a paper for a conference whose deadline is approaching. He switches on his computer, and he checks his email. He spends some time replying the messages, checks a couple of entertaining websites, and it is time for lunch; in the afternoon, he remembers that he had to have some exams graded for the week after, and does it. The following days are pretty much like this one; the deadline is approaching and Joe has not even started writing his abstract. Is this a case of weakness of will? Insofar intentionally grading exams does not prevent Joe from applying to the conference, he is not holding conflicting intentional states. However, as the deadline approaches, there can be certain intentional actions that prevent Joe from applying for the conference, because he will not have time to write the abstract. Therefore, a procrastinator is also weak-willed when he intentionally does something that prevents him from achieving a previous goal, and nonetheless she does not abandon her intention. For example, Joe cannot intend to apply for twenty conferences on the same week: there are not enough hours in a day in order to write so many abstracts. Or, suppose that a more important task comes out, and Joe has to drop his intention to apply to the conference. Those are not cases of weakness of will – sometimes, we would like to achieve different goals that, unluckily for us, are not

101 See, for a recent collection of essays on this topic, Andreou and White (2010).

compatible amongst them. Thus, procrastination also entails weakness of will when the agent never drops her intention to achieve her end, but intentionally does other things that, eventually, enter into conflict with her intending the means.

Chapter 2. The normativity of practical commitments

Practical reasoning is guided by regulative principles: an agent infers what she should do in the light of what reasons for action she has. Intentions have a normative dimension, as long as they are required to be consistent amongst them, and with the normative judgements concluding practical reasoning. However, this claim needs further development:

> The claim that "the intentional is normative" is the claim that any adequate account of the nature of intentional mental states must employ normative terms (or at least must mention the properties and relations that these normative terms stand for). But different versions of this claim will give very different accounts of the exact role that normative terms must play in adequate accounts of the nature of intentional mental states. [1]

The claim that intentionality is normative is not necessarily essentialist. I am not concerned with "shmagents"[2], this is, with subjects who are indifferent towards the constitutive standards of agency, and therefore are unmoved by them. Even if it is accepted that it is possible to be a "shmagent"[3], this possibility does not threat the claim that reasons, evaluative rules and normative requirements are constitutive of agency insofar agents engage in practical reasoning. Rational norms, or requirements, do not only play a role in practical reasoning, guiding behaviour through their justificatory capacity—they are also necessary for making sense of the others' actions. We assume that other agents behave rationally, and when they do not, we look for an explanation of their deviation from the normative standards. I believe that these requirements are

1 Wedgwood (2007a: 85–86).
2 Enoch (2006).
3 for an argument against, see Ferrero (2009).

evolutionarily shaped and socially transmitted, and so does the normative dimension of agency. Besides their function as enablers of inter and intra-personal coordination[4], sense-making is also a further function of normative requirements.

In this Chapter, my aim is to explore three normative aspects of practical commitments, on the one hand, and to suggest an account of the requirements governing rational agency, on the other. §2.1 is concerned with the first of these two aims: the analysis of three normative features of practical commitments. The first of them concerns the link between reasons and normative judgements—this is, the normative structure of practical reasoning. I will argue that (i) reasons are facts, and not mental states; and (ii) they have to be *possessed* by that agent: effective reasons are subjective (or agent-relative) reasons. Second, I will defend that the conclusion of practical reasoning is a normative belief, and not an intention. Finally, I will discuss the 'bootstrapping objection', proposed by Bratman[5]. The objection can be stated as follows: "you cannot bootstrap a reason into existence from nowhere, just by a forming an intention"[6].

Section 2.2 is devoted to the analysis of normative requirements. Attributions of irrationality are made on the basis of a violation of some of those requirements. I will first present a recent debate about the appropriate formulation of normative requirements: wide versus narrow-scope formulations (§2.2.1). I will argue that narrow-scoped requirements have the advantage of gathering the directionality and agent-relativity of practical rationality. In §2.2.2, I will suggest an alternative formulation of three rational requirements: enkrasia, resolve, and means-ends reasoning, whose violation is the basis for attributing akrasia and weakness of will (see §1.2.3). I will defend that enkrasia is better understood as a restriction, rather than a requirement, and that the means-end coherence requirement is derived from a more general rationality principle regarding consistency amongst intentional states, which I call *resolve*.

4 Bratman (2009b).
5 Bratman (1987).
6 Broome (2001a: 98).

2.1. Reasons, intentions and practical reasoning

Practical commitments have a normative dimension, which goes through different elements of agency. In this Section, I will explore three of these elements: reasons, normative judgements, and the relation between reasons and intentions. These three topics are very controversial in the literature. In a very broad sense, they reflect the problematic relation between desires and oughts—between what the agent wants to do, and what she ought to do. The main problem is that oughts do not motivate, and desires do not justify. Insofar rational agency displays these two features, motivation and justification, to stablish the appropriate relation between oughts and desires is central to the analyses of rational agency.

2.1.1. Reasons for action

Agents may have reasons for performing an action, for holding a belief, or even for feeling in a particular way[7]. For example, Sarah has reasons to believe that the Earth is not flat. Also, the fact that Sarah works tomorrow is a reason for her to set the alarm clock at 8 am, and furthermore, it is the reason why she has done so. Having told a lie is a reason to feel ashamed, and so on. In a very general and broad sense, a reason is the answer to the question "why?", or what "counts in favour of" doing, believing, or feeling something[8]. In this dissertation, I am only examining reasons for action, this is, those answering the question "why will/should/did agent A φ?".

There are two main kinds of reasons addressed in the literature: normative (or justificatory) and motivating, sometimes called explanatory[9]. The main idea behind this distinction is that explaining an action and justifying it are two different tasks. Explaining an action would consist in making the agent's motivations explicit, or addressing to its

7 See Skorupski (2002).

8 This definition can be found in Scanlon (1998).

9 This distinction is widespread; see for instance Raz (1975); Williams (1982); Smith (1987); Parfit and Broome (1997); Dancy (2000); Audi (2001); Schueler (2003); Finlay (2006).

causes; and justifying it would require to give a reason that turns the action correct or right (not necessarily *morally* right). So, normative reasons have features which motivating reasons lack: for instance, a normative reason can be a good or a bad reason, while a motivating reason cannot[10], and it is possible to feel motivated to φ and at the same time not being able to justify to φ.

There are two main positions in the debate on the ontology of reasons[11]. On the one hand, some authors argue that reasons are *mental states*, such as desires, intentions or beliefs; after all, motivating reasons are supposed to have a causal role in the agent's actions[12], and thus it makes sense that reasons are somehow connected to mental states. Following the Humean account of action, propositional *contents* are not able to "trigger" action, but desire-related mental states are. A Humean theory of action would claim that the agent's desires[13] cause the action, and therefore those desires are the reasons why the agent performed the action. If we aim to attribute reasons a causal role in the explanation of action, it seems natural to identify them with mental states. Beliefs are, traditionally, denied to have motivational force[14]; thus, reasons have to be reducible to, or accompanied by, motivational attitudes such as beliefs[15]. Internalism about reasons, thus, requires that there is a connection between motivation and reasons.

On the other hand, it is argued that reasons have normative force: they allow us to judge an action as correct, justified, required, etc. Mental states, however, do not seem to be able to do this –facts do. Thus, some reasons may not belong to the agent's set of mental states and still be reasons for that agent to φ. Externalists claim that there is no conceptual link between having a reason and being motivated accordingly, because reasons

10 Although this claim is frequently endorsed, a defence at length can be found in Schueler (1995).

11 See, for an overview of the problem, Alvarez (2005); Everson (2009).

12 Davidson (1963); Smith (1995).

13 It is important to note that mental states are sometimes confused with both the fact that an agent is holding a mental state, or with the content of the mental state itself. For instance, it is easy to mix up the fact that Sarah desires to buy a new car, with the mental state corresponding to Sarah's desire (whose content would be to buy a new car), and also with the fact that Sarah buys a new car, which is the content of her desire.

14 I do not believe this is necessarily the case; for example, my belief that I will succeed at performing certain task can motivate me to do it. I am just pointing out a classic assumption for the Humean theory of action.

15 See Davidson (1963); Williams (1982); Mele (2003a).

are facts, and not mental states[16]. External reasons do not necessarily motivate behaviour, but have normative force[17]. Externalism distinguishes between good and bad reasons, and not only strong and weak reasons. However, this view leaves room for a strange situation, in which an agent acts for no reason (although she believes that she has reasons). Parfit[18] claims that reasons are provided by facts, and hence it is possible to act for no reason, if we are mistaken about the fact we take as such. However, this does not turn our behaviour irrational: it is rational to act according to our beliefs, regardless of whether they are true or false beliefs.

I will defend an externalist and perspectivist account of reasons. On the one hand, I will argue that reasons are facts, and that they need not to be necessarily accompanied by a motivational attitude. However, I do not endorse the view that reasons are agent-independent. I will argue that agent-neutral reasons express norms that aim to be universal.

Reasons are facts

One and the same reason can be used to explain why an agent *has performed* an action and why she *should perform* an action (the same action, or a different one). For example, the fact that Sarah is seeking revenge is Sarah's reason to kill her aunt; but this same reason would justify that Sarah visits a psychiatrist in order to have her aggressive impulses revised. But, how could a mental state make another fact appropriate or correct (such as visiting a doctor)? Álvarez[19] argues that mental states do not haver normative force; and it would be highly implausible that one and the same reason changes its ontological status depending on what role it plays (motivating or normative). Therefore, she concludes, all reasons are facts[20]. The normative power of justificatory reasons would be similar to the

16 See Stout (2004); Korsgaard and O'Neill (1996); Setiya (2004); Setiya (2007b); Raz (2009).
17 Broome (2004); Dancy (2004b).
18 Parfit (2001).
19 Alvarez (2010a).
20 Ibid., 49.

normative power of evidence: they raise the probability of another fact to be the case[21], making it appropriate to believe or to act. Thus, generally speaking, normative or justificatory reasons are used in practical reasoning about why an action φ should be performed. Explanatory reasons are supposed to be used for making inferences that allow the agent to understand why something is the case: an action which has been performed, or an action which will be performed[22]. Motivating reasons can be used to explain why an action was performed; there can be explanatory reasons that the agent is not aware of, and thus do not motivate her at all. For example, if Sarah suffers from an Obsessive Compulsive Disorder (OCD), the fact that her hands may be full of harmful micro-organisms is the reason that motivates her action, but from our point of view, her action is better explained presenting her OCD condition as a reason.

A traditional argument against the claim that all reasons are facts is that reasons seem to have causal power: they cause the agent to intend to perform an action. It is assumed that the conclusion of practical reasoning is an intention (or an action), causally related to the reasons to perform it. I think this account is misleading, and I will argue against it in §2.1.2. Assessing what reasons for performing an action an agent has and deciding what to do in the light of the previous assessment are two different tasks, and are subject to different rules. Also, there is a link between the normative belief produced by practical deliberation and the formation of the intention. This link takes the form of the enkratic rational requirement (see §2.2.2). Hence, as long as reasons do not "produce" intentions, but normative beliefs, they need not to be accompanied by motivational

21 See Kearns and Star (2009). I am aware that the evidentialist approach to reasons entails some difficulties, specially concerning the problem of how to raise the probability of a normative fact to be the case. For instance, the fact that someone is in trouble raises the probability of the normative fact "Someone should help this person", and is a normative reason to help her. Explanatory reasons are evidence-like, because they actually serve as evidence in comparing and assessing beliefs and belief's degrees of certainty. But the role of normative reasons in increasing the probability of normative facts is far from clear.

22 It is quite plausible that explanatory reasons are, in fact, evidence for belief. If I believe that a fact F can explain another fact F', then F is evidence for F': it makes F' more likely to happen, or to be true. I will not, however, analyse reasons for belief in this dissertation, although I acknowledge that their connection with normative reasons for accepting a proposition (which is an intentional action, contrary to belief) is highly relevant to the discussion on the connection between practical and doxastic inferences.

attitudes (although, of course, the agent can have whatever motivational attitude towards the facts they take as reasons).

Reasons are perspective-dependent

The distinction between objective (or agent-neutral) and subjective (or agent-relative) reasons was stated by Nagel[23], and has been recently developed by Schroeder[24]. He sets the distinction between subjective and objective reasons to believe in the following way:

> If Max is smiling, that is reason to believe that he is happy. But if no one realizes that Max is smiling, no one has that reason to believe that Max is happy. I'll call the sense in which the fact that Max is smiling is a reason to believe that he is happy, even if no one knows about it, the objective sense of 'reason', and I'll call the sense in which in this case no one has a reason to believe that Max is happy the subjective sense of reason. [25]

Schroeder argues that reasons are not correctly thought of as being objects in the world that are reasons by themselves. He argues that the traditional view on reasons, the "Factoring Account", implies that there are things "out there" in the world that are reasons (for example, there is a reason to φ), and sometimes we have them (thus, we have a reason to φ), in the same way we own a ticket to the opera. However, this account cannot explain what is wrong with some special cases, for instance those in which the agent has a false belief, which she takes as her reason to φ. To illustrate this point with Schroeder's example (based on William's): imagine that Bernie is in a bar and asks for a gin and tonic. Then, by mistake, the waiter puts gasoline in his glass (instead of the ordered drink). What would be Bernie's reason to take a sip from that glass? After analysing several candidates for being a reason, and for being Bernie's reason, Schroeder concludes that Bernie's reason to take a sip is the fact that the glass contains gin and tonic: this seems counterintuitive, so long as the glass actually contains gasoline Schroeder argues that Bernie's reason is

23 Nagel (1970).

24 Schroeder (2007); (2008); (2009).

25 Schroeder (2010: 1). I am fully aware that reasons for belief differ in some (crucial) aspects from reasons for action; however, I have chosen to illustrate the difference through this quotation because (i) the difference between objective and subjective reasons is the same regardless of whether it is applied to reasons for belief or to reasons for action; and (ii) it gathers in a clear and precise way what this difference actually is meant to differentiate.

subjective. A fully informed witness can reasonably argue that Bernie does not really have a reason to drink from that glass; this would correspond to an objective use of reasons.

While subjective reasons for action are used as premises in practical reasoning, objective reasons designate norms and evaluative principles, which we aim to apply not only to our own reasoning, but as a generalized norm. For example, we might say that there is no reason to set a cat on fire, or to attack a Chelsea fan. However, a sadist *has a reason* to set a cat on fire ("it is fun"), and a West Ham football club hooligan *has a reason* to attack a Chelsea fan ("they deserve it", or even "it is fun"). The first sense of reason would be objective, following Schroeder, and the second would be subjective. But we would want to say that subjective reasons (at least those reasons) are not reasons at all, or at least, that they are not *good* reasons, independently of the motivational relation they stand in with the fact that is offered as reasons. Denying the truth of "having fun is a reason to set a cat on fire" is the same as stating that there is no norm that allows anyone to take the fact of having fun as a reason to make someone suffer. Thus, it refers to a norm, specifically, the inferential norm that allows the agent to infer "I should set a cat on fire" from the premise "setting a cat on fire is fun". It refers to evaluative principles insofar I can translate "nobody has a reason to set a cat on fire" into "nobody should infer from any fact (such as a perspective of having fun) that she should set a cat on fire". Objective reasons, then, refer to what inference rules are legit, good, or valid. If I assert that the fact that Max (from the example above) is smiling is an objective reason to believe that he is happy, I am not offering any reason to believe that he is now happy: I am just claiming the validity of an inferential rule.

Similarly, reasons based on false beliefs face the same duality regarding objective and subjective reasons. If my reason for running is that a bear is chasing me (or so I believe), but the bear turns to be a friend dressed as bear, then, in a sense, I do not have any reason for running, for the fact I take as reason is not really a fact. But the sense in which I do not have any reason for running is an objective sense of reasons, and I believe

that it does not refer to reasons for action, but to norms of inference and evaluative principles.

The distinction I want to draw between the objective and the subjective sense of reasons can be summarized as follows. The validity of the rule "being chased by a bear is a reason to run from it, as long as you want to be alive" (which would be equivalent to the conditional command "if you are being chased by a bear and want to live, run from it!") makes "I am being chased by a bear" a reason that favours the conclusion "I should run from it" (given that I want to survive). But the rule is not a reason itself, although it corresponds to the objective use of "reason". When it is said that there are reasons to run from bears (objectively), the function of this assertion is only to make the fact that I am being chased a reason. In this sense, subjective reasons need from objective reasons. But I believe that the conceptualization is misleading. Reasons to act, as something that make a conclusion valid, this is, as premises that entail a normative conclusion, are always subjective. But, in order to use a fact as a reason, the agent needs to use some norm or inference rule, which enables drawing the conclusion. Thus, the concept of reason I will use corresponds to the subjective sense; I will refer to inference norms, or rules, to refer to "objective reasons". Of course, agents may believe that the rule or norm that they apply in order to use the fact of having made a promise into a reason for doing what is promised is true, valid, or correct. This is: in order to present a fact as a reason, a norm (objective reason) has to be applied. In what sense is it an objective norm, then? I believe that in none. Norms are as agent-dependent as the reasons enabled by these norms.

To sum up, I have argued that objective reasons are not reasons, but inference rules. For example, "having made a promise is a reason to do what is promised" would correspond to the objective sense of reasons. My claim is that this is not a reason at all: it is a norm, a rule, that allows to take the fact "I have made a promise" as a (subjective) reason to do what I have promised to do. This would be the subjective sense of reasons. As long as I take both to be agent-relative (reasons and norms), I believe the distinction objective / subjective is misleading.

I have argued so far that reasons are facts, on the one hand, and that reasons for action are subjective reasons: in order to be used as a reason, a fact has to stand in a particular relation with the agent who uses it. I cannot use as a reason something that I believe is not the case[26]. But this does not entail that reasons *are* mental states. In order to present a fact (to others) as an argument for doing something, that fact needs to be the content of some mental state of mine –otherwise I would not be able to offer it (neither as a reason for nor against) as a part of an argument. It does not follow, however, that I am presenting my mental state as a reason for or against doing something. The same goes for reasons for action used in deliberation: if I am using those reasons, they have to be the content of some mental state of mine, insofar I am deliberating. The reasons I am using are facts, something I present to others and to myself as true. Of course, I can be wrong about the facts I use, but this does not mean I do not have any reason for action. A different issue, thus, is whether I am justified—from the point of view of some external, objective standards—in using certain facts as reasons for action or not: this is what objective reasons refer to.

Reasons and oughts

Normative reasons, then, are facts presented by an agent in order to justify an action. Practical reasoning consists in evaluating and assessing one's reasons regarding one or more possible alternative paths of action[27]. When evaluating one's reasons, an evaluative judgement obtains, which can take the form of an ought. It is not necessary, though, that the practical judgement that concludes practical reasoning takes the form "I (or someone

26 This does not entail that I have to know the fact, or even believe that it is true; I can just hope it is. For example, I hope that my father will be happy if I buy him a present for his birthday, and this is a reason for me to buy the present; however, I am not sure that he will be happy, because we had a recent argument and he might feel angry at me.

27 Of course, it is possible (although not very frequent) to wonder "What ought I do?", and then look for alternatives, and find supportive reasons. But, in general, practical reasoning concerns the evaluation of alternatives; one may also wonder "What *can* I do?", but I would probably consider this question an instance of theoretical reasoning, for the question is not about justifiability, but availability.

else) ought to φ". It can take different forms. Following Thomson[28], there are two classes of normative judgements: evaluatives and directives. The former class includes evaluations of a certain situation and/or its properties; for instance, to assert that smoking is bad for health, that a carving knife is a well-made or a defective carving knife, or that this is a better guide dog than this other one. Directives, on the other hand, are meant to express what something or someone ought (or should) to do or to be: a guide dog ought to stop at the pedestrian crossing, or people should aim to be happy. If practical reasoning aims to decide what to do, it seems that its conclusion should be a directive judgement. Although I will use hereafter directives as examples of normative judgements, evaluatives can also conclude practical reasoning. For instance, I can conclude that, between two options, one is better than the other. The directive judgement "I ought to choose option A" follows from "Option A is better that option B" and "I ought to choose the best option", but these intermediary steps are not necessary for choosing option A rationally. What is indeed necessary is that the agent evaluates her alternatives, given the reasons she has; thus, the conclusion of practical reasoning can be either a directive or evaluative judgement. In general, an agent-relative ought-proposition is such whose falseness entails the violation of a rule, which is an agent-neutral ought-proposition. For instance, if Bert ought to do what he has promised to do, and he does not, then the norm "promises ought to be kept" is violated. This would be equivalent to "correct promises are fulfilled promises". Other modal modifiers, such as *must*, or *should*, have the same meaning that *ought* in this context.

Normative reasons are the basis for both evaluatives and directives. If I judge that alternative A is better than alternative B, I judge so because of some reasons, that justify the superiority of A over B. Evaluating these reasons and comparing alternative paths of action is the central function of practical reasoning. As I will argue in the next Section, the conclusion of practical reasoning is a normative judgement, which of course can guide intention formation, although the formation of this intention is not a part of practical

28 Thomson (2007: 240).

reasoning, but a part of rational, or practical, agency. However, is there a difference between judging that one has more (or stronger) reasons to do A than to do B, and judging that one ought do A? My answer here would be that, even of there is a difference, it can be disregarded. There might be some cases in which an agent agrees in that, when facing two alternatives A and B, alternative A is better than B. And yet fail to judge that she ought to A. However, I cannot imagine a case in which this agent, if asked whether someone willing to choose between A and B, does not judge that the chooser (whoever it is) ought to choose A. To put it differently: no examples of agents judging that A is normatively better than B, and at the same time, judging that the ought not to do A, come to my mind. It is possible that the agent has doubts regarding the weight of her reasons, or that the agent concludes that two alternative reasons are equally justified. And of course it is possible not to reach a conclusion. My point is that practical reasoning is relevant for agency insofar the agent reaches a conclusion, even if it is provisional, or sensitive to new evidence. When it is argued that practical commitments entail a normative dimension it is because of the role of normative judgements in practical agency. Thus, in the next Section I will argue that a normative judgement (either a directive or an evaluative) is the conclusion of practical reasoning, but it is also possible that an agent engages in practical reasoning without reaching any conclusion; I will leave these cases aside, for they are not central to rational agency. Similarly, I will discuss the normative requirements of rationality in §2.2. Normative requirements are the rationality norms or rules whose violation constitutes irrationality—*akrasia* would be a violation of the *enkratic* requirement, and *weakness of will* violates the *resolve* requirement. I believe that a requirement that demands the formation of a normative judgement after the evaluation of reasons is unnecessary. It could go as follows: "An agent ought to judge that she should[29] do what she believes she has most normative reasons to do". I believe this requirement would be a truism. Compare to *enkrasia*: following a standard formulation, which I will

29 I am using ought and should interchangeably here; I use two concepts to make the meaning of the sentence clearer.

discuss in §2.2.2, it states that an agent ought to intend to do what she judges she ought to do. This requirement is relevant insofar it is possible, and even frequent to some point, that people do things that they judge they ought not to do—think of procrastination, for instance. On the other hand, it is not frequent, and I doubt whether it is possible, to judge that one has more and stronger reasons to A and, at the same time, judges that she ought not to A. Therefore, in what follows, I will assume that every agent that reaches an evaluative judgement can also be attributed a directive judgement. Both of them can conclude practical reasoning, and both of them have normative guidance.

2.1.2. Judging and intending

Broadly speaking, deliberation is a mental activity consisting in the examination, evaluation, and assessment of reasons for action. The conclusion of deliberation is a judgement about what we should be done, given certain reasons. The steps from reasons to normative judgements are inferential. Together, deliberation and judgement make up practical reasoning. The transition from reasons to judgements is not volitional, but merely normative. Our reasons normatively commit us to coming to believe that we should do something: this normative commitment expresses the normative requirements of practical reasoning.

Practical reasoning contrasts with theoretical reasoning: while the former is directed towards action, the latter aims to elucidate how the facts stand. This starting point has led the majority of philosophers to claim that the conclusion of practical reasoning is an intention[30], or an action[31], or any of them –decisions or actions [32]. Following Audi[33], I will defend that the conclusion of practical reasoning is a belief, and not an intention, neither an action:

30 See Brandom (1998); Broome (2002); Stroud (2003).

31 See Dancy (2004a); Tenenbaum (2007).

32 Alvarez (2010b). See Streumer (2010) for a recent overview on the debate about the conclusion of practical reasoning.

33 Audi (2006).

> We must distinguish between the conclusion of a practical argument, which I take to be a proposition, and what corresponds to it in [the subject]'s reasoning: concluding that reasoning, by inferring the conclusion from the premises. Typically, the conclusion will be the kind of proposition we think of as a practical judgment, and the concluding of the reasoning with that judgment will be an instance of judging that the action in question is, say, the thing to do. [34]

I will now provide three arguments to support this thesis: (i) the possibility of reasoning about what someone else should do, (ii) the completeness of the process of practical reasoning without forming an intention, and (iii) the ambiguous character of hypothetical or exploratory deliberation.

Advice and second-person practical reasoning

First, although practical reasoning is directed towards action, it is not necessarily *one's own* action; I can deliberate about what you should do, and judge that you ought to φ. Theories defending that the conclusion of practical reasoning is an intention, or an action, consider this form of reasoning *theoretical*. I find this misleading, for the kind of inferences that one makes when deliberating about what one should do are essentially the same to those made when deliberating about what a team, or someone else, should do. It can be argued that one has no agential authority over someone else's actions; this is, that I can deliberate about what you should do given my policies, deliberative commitments, values and so on. But this poses no problem to the thesis that the conclusion of practical reasoning is a normative belief. The difference lies in that the conclusion of my reasoning exerts no control over your choices, nor you are rationally required to form the intention to do what I judge you should do. These two bonds (control and normative requirements) between deliberation and choice do not belong to practical reasoning, but to a broader model of rational agency, as I suggested in §1.2.

It can be argued that, when reasoning about what someone else should do, we are in fact engaging in a theoretical reasoning process. For example, Álvarez[35] argues that

34 Ibid., 68.

35 Alvarez (2010b).

"practical reasoning presupposes a goal in the person who engages in the reasoning, which is precisely the thing wanted and what gives the point of the reasoning and of the action to which the reasoning leads "; however, she adds: "Unless, that is, one is just reflecting on how practical reasoning works, or reasoning on someone else's behalf, as a detective might when trying to guess how someone might have acted"[36]. I believe that these two exceptions (exploratory reasoning and second-person reasoning) make it difficult to require practical reasoning to be driven by an agent's goal. I agree in that practical reasoning is usually prompted in situations in which the agent faces a choice, or intends a goal; my point is that this is not necessarily the case, because being driven by a goal is not what characterizes practical reasoning.

The conclusion of practical reasoning does not aim to describe how things stand; it rather aims to justify an action, a goal, a value, etc. Second-person (or third-person) practical reasoning is prompted, amongst other situations, by situations in which an agent asks for advice, or that the deliberative agent wants to give some advice to someone else. Imagine, for instance, that a father advises his son: "Son, you should study law rather than philosophy; I have evaluated all the reasons you gave me, even under the lights of your own standards, but I do not see how they overcome the hunger you will experience as a philosopher". The reasoning this father has done regarding his son's academic future is very similar to the one he would have done when assessing whether to study philosophy himself. The difference between first and second-personal practical reasoning lies in (1) the amount of information available for each agent, and (2) the evaluation mechanisms used to assess the reasons for and against studying philosophy or law. Although, as Andreou[37] argues, judgements about what an agent ought to do can be made from within the deliberative agent's (the adviser) standards and values, or within the advisee standards and values; thus, the difference stated in (2) is not always the case.

Complete practical reasoning does not require intention

36 Ibid., 367.

37 Andreou (2006).

Second, practical reasoning can be complete without the formation of an intention. If its conclusion was an intention, all exercises of practical reasoning that do not end up in the formation of an intention would be incomplete, besides normatively incorrect, and volitionally flawed. However, cases of practical reasoning that stop in a judgement but do not form the intention seem to be complete processes reasoning. The following example is provided by Mele:

> Consider Joe, a smoker. On New Year's Eve, he is contemplating kicking the habit. Faced with the practical question what to do about his smoking, Joe is deliberating about what it would be best to do about this. It is clear to him that it would be best to quit smoking at some point, but as yet he is unsure whether it would be best to quit soon. Joe is under a lot of stress, and he worries that quitting smoking might drive him over the edge. Eventually, he decides that it would be best to quit – permanently, of course – by midnight. Joe's cognitive decision settles an evaluative question. But Joe is not yet settled on quitting. He tells his partner, Jill, that it is now clear to him that it would be best to stop smoking, beginning tonight. She asks, "So is that your New Year's resolution?" Joe sincerely replies, "Not yet; the next hurdle is to decide to quit. If I can do that, I'll have a decent chance of kicking the habit". [38]

Joe might be weak-willed, if he intended to lead a healthier life, and, at the same time, delays the decision to quit smoking, as I argued in §1.2.3. He may even be akratic if, having judged that he ought to quit, he forms the intention of not quitting, and still holds his judgement (see §2.2.2). But Joe's practical reasoning process is complete: he has evaluated his reasons, and decided *that* the best thing for him would be to quit smoking – although he has not decided *to* quit smoking yet. Hence, Joe may be irrational from the point of view of practical agency, but not because his reasoning is neither incorrect, nor incomplete. There are cases in which an agent judges that a given option is the best one amongst all the available alternatives, but she is not facing the moment of choice and prefers not to commit herself to a path of action. Suppose that Ingmar has run certain medical tests in order to know whether he suffers from cancer. He receives his analysis in a closed envelope on, let's say, November 1st. He decides not to open the envelope at home, but at the doctor's surgery; the appointment is on December 1st. However, Ingmar deliberates at that moment about what he should do if the results are positive for cancer.

38 Mele (2003a: 199).

He judges that, if the cancer tests are positive, he should undergo whatever treatment the doctor proposes, and if, eventually, the cancer develops to a terminal phase, he ought opt for euthanasia. Does this judgements commit Ingmar to form a conditional intention? I believe not; he has just decided to postpone the act of choice because he is not required by the circumstances to choose at the moment of receiving the envelope; however, his practical reasoning process is complete.

Against this argument, it could be claimed that, had the agent completed her reasoning process (this is, that she has reached a conclusion), then we would have to explain why Joe, or Ingmar, are behaving irrationally (and, if they is not, then their reasoning would be *theoretical*). My response would be that practical reasoning and rational agency are indeed related, but they do not refer to the same concept. Failing to act as intended can be a failure of *rational agency*, or practical agency, but I do not believe this to be a failure of *practical reasoning*. Insofar I acknowledge a rationality failure between those steps, the only disagreement would be whether this failure belongs to the sphere of *rationality* or to the sphere of *reasoning*, but this does not affect the claim that the conclusion of practical reasoning is a normative belief.

Hypothetical or exploratory practical reasoning

The difference between hypothetical (or conditional) and exploratory deliberation can be a matter of degrees of probability. I sometimes wonder what I would do, what I should do, given certain hypothetical scenario. Many times, practical reasoning can lead to the formation of a conditional intention[39]: "If my sister wants to, I'll go to the cinema with her; otherwise, I'll stay home and read a book". Here, I reason about what I should do in a conditional scenario, and I form an intention now to perform that action if the conditions

39 In fact, most intentions are subject to implicit *ceteris paribus* conditions Klass (2009). It is quite unusual to intend to φ no matter what. For example, I have decided to visit my mother tomorrow; but this intention can be legitimately revoked by many conditions, such as breaking my leg, having a fever, a fire in my house, a friend having suffered a car accident... This is, unless I come up with stronger reasons to revoke my intentions. In this sense, as Bratman points out, the inertia or stability of intentions consist in the absence of reconsiderations. But, of course, it is not irrational to reconsider an intention in the light of new reasons for action.

apply. However, suppose that I deliberate in the following way: "What would I do if my building was being overrun by zombies? I think that, given my poor shooting skills, and my actual lack of guns, I should try to escape through the flat roof". This is a piece of practical reasoning which I take to be correct. However, I do not believe that I am required to form the conditional intention to escape through my flat roof in case of a zombie invasion. Conditional intentions entail volitional commitments; however, they delegate the control of the goal (which, in this example, would be to survive a zombie attack) to "anticipated situational cues, which (when actually encountered) elicit these responses automatically" [40]. Also, conditional intentions regulate the formation of further intentions. If I intend to go to the cinema with my sister (in case she wants to), then it is contradictory that I unconditionally intend to go to the lake the same day. However, in the zombie scenario, my belief about what I should do in case of being threatened by a bunch of living dead does not condition nor control my other intentions. I keep closing my windows in a way that it would difficult that I escape through them, without holding conflicting intentions.

The scope of possible scenarios between my sister wanting to watch a film and me being attacked by zombies ranges from being quite likely to happen, to almost impossible. Hypothetical reasoning is thus very similar to exploratory reasoning: we form a (conditional) intention only insofar an intentional response is required –for example, because we intend to do something. I take to be both examples valid processes of practical reasoning; thus, the formation of an intention follows deliberation and judgement only when they are used as means to choose, but not because of their normative structure, or because of a motivational relation between the reasons used as premises and the intention formed thereafter.

To sum up, it is widely accepted that practical reasoning is directed towards decision or action, while theoretical reasoning aims to form a belief about how facts stand. I endorse this differentiation; however, I do not agree in that practical reasoning

40 Gollwitzer (1999).

necessarily prompts or requires the formation of an intention, or the performance of an action. Commonly, judgements about what one should do support, and are the basis for, forming an intention in accordance with that judgement. This bond has both a normative and a volitional dimension. Normatively, an agent is subject to the enkratic requirement – which, as I will defend in §2.2.2, requires from the agent that she does not intentionally do what she judges she ought not to do, rather than requiring to intend to do what she judges as best. The formation of an intentional state, or the performance of an intentional action, is required only when the agent holds a previous and active goal to which she is committed. But she is not required, for example, to intend the means if she intends the end because intending that end provides new reasons for action; this would be a case of bootstrapping.

2.1.3. The bootstrapping objection

An intuitive view about means-end reasoning is that, if you intend an end (for example, to cook pasta), you have reasons to intend the means (to buy pasta), or at least, you have one reason that you did not have before intending the end:

> That [an agent] had committed herself to doing something, or resolved to do it, would be for her a new reason for doing it, much in the way in which promises to others provide most people with new reasons for doing what they have promised to do. This is at least part of the way in which advance decisions work for most people. [41]

The nature of reasons has drawn much more attention than their creation process. In a broad sense, we can choose what to intend and what to intentionally do. This intuition would seem to suggest that we can choose, at least, our motivational reasons, or at least weight and compare them voluntarily. Searle (2001) is the most representative defendant of what Watson[42] calls the *Autonomy Thesis*. It states that an agent can freely create a normative reason for action, insofar she can freely undertake an obligation; this obligation

41 Sobel (1994: 249).

42 Watson (2009).

is a desire-independent (following Searle) reason for action, namely, a reason for performing the obliged action:

> The presupposition of the freedom of the agent is crucial to the case as I have described it. From the first-person point of view, by freely undertaking to create a reason for myself, I have already manifested a desire that such and such be a reason for me. I have already bound my will in the future through the free exercise of my will in the present. In the end all these questions must have trivial answers. Why is it a reason? Because I created it as a reason. Why is it a reason for me? Because I have freely created it as a reason for me. [43]

This seems an overestimation the power of the will and an overgeneralization of normative powers. The Autonomy Thesis enables the possibility of making self-directed commands: I command myself to do something, and through this act I give myself a new reason for action[44]. Commands (and promises) provide the agent a reason for doing as commanded, in the absence of any other reasons, as I will expound in Chapter 4. Considering intention as a form of self-command enables the possibility of unlimitedly creating (normative) reasons. A further problem of the self-directed command view is that the agent has both the authority to issue and to revoke commands, thus it is not clear in what sense the agent is bound by the command she has issued to herself[45]. Hence, the capacity to impose obligations to ourselves, or to freely create reasons for action, seems to be trapped in a vicious circle.

It is then plausible that the process of reason creation is subject to some limits. For instance, Raz[46] argues that it is possible to create reasons, but this capacity is constrained by their normative nature. A connection between the reasons created and pre-existent ones is needed:

> We cannot create reasons just by intending to do so and expressing that intention in an action. Reasons precede the will. Though the latter can, within limits, create reasons, it can only do so when there is a non-will based reason why it should. [47]

43 Searle (2001: 189).
44 See Velleman (1989: 99) for a defence of a similar view.
45 See Ferrero (2006).
46 Raz (1986).
47 Ibid., 84.

Hence, there are two confronting perspectives. The first claims that, as far as an agent is able to commit herself to the performance of an action, this commitment represents a reason for action, which has been created by the agent. The other expresses the implicit normative limits of reason creation: it does not seem possible to freely create reasons because, if this was the case, then the normative character of reasons would be confined to the boundaries of the will.

To illustrate the problems of creating reasons from the perspective of rationality and normativity, let's imagine the following situation[48]: Sally is walking down the street when a cent accidentally falls from her pocket to a really dirty puddle. After realizing her slip, she thinks that picking a cent is not a strong reason to get her hands dirty. Then, she takes a Euro from her wallet, drops it into the floor, and thinks "well, now I have enough reasons to get my hands dirty". She picks her 1,01 Euros, and keeps walking.

Is there anything wrong with Sally's reasoning? In fact, there is nothing wrong. This story is told as a joke because it implicitly expresses that Sally (or any other stingy person) was indeed motivated to pick up the cent, but then she would be revealing her own evaluation mechanisms –in this case, her stinginess. She has chosen a path of action first, and then has acted to create normative reasons that justify it.

It seems that having the intention to pick up the coin was a sufficient reason for Sally to get her hands dirty. However, this assumption would leave room for having at least one reason to perform whatever we intend. This consequence seems counterintuitive, and thus the possibility of taking our intentions as reasons is controversial.

If we concede intentional states the capability to create new reasons for action, then it is possible to encounter awkward situations. Broome illustrates this point through the following scenario:

> Suppose you are wondering whether to visit Paris, but have not yet made up your mind. There are reasons in favor and reasons against. Whether or not you ought to go depends on the balance of reasons. Now suppose you make up your mind to go, so now you intend to go to Paris. Ought you to go or not, now? What does that now depend on? [49]

48 This example is actually a joke, whose target, instead of Sally, is the stereotype of a scrooge person.

49 Broome (2001a: 98).

In his analysis of intentional action, Bratman (1987) claimed that, if an agent has no reason to φ, then intending to φ do not create one[50]: this is what the bootstrapping objection claims. Broome uses it for arguing that, besides reasons and oughts, there is a conceptual need for a wide-scoped concept of normative relation, what he calls a normative requirement. So, intending to φ normatively requires to φ, but it does not provide a reason for doing so. If John were asked why he is going to return Sarah's book, it seems that answering "because I intend to" is a rather uninformative explanation. It does not seem that his answer provides a satisfactory justification on why he should return Sarah's book. So far, we accept that the bootstrapping objection is correct. The bootstrapping objection also applies to means-end reasoning. Having an intention and knowing the means does not imply that one has a reason to intend or perform the means, because it it possible that one has no reason (in a subjective sense) for performing the end in the first place. The bootstrapping objection claims that if an agent has no reason to φ, she cannot create a reason to φ just by intending to φ. This is, if the agent has no available facts to use as reasons to φ, then by changing our mental state towards it will not create a reason. Mental states, I have argued, are not reasons. The same objection can be found in theoretical reasoning. Let's imagine that Sarah believes that p. Then, she can infer that she believes that she believes that p. But the fact that Sarah believes that p is the case is not a new reason that supports/justifies/explains why she believes that p, nor is evidence for p. The same way, if Sarah has no available reason to φ, she cannot create one by wishing that she desired to φ. To sum up, a mental state whose content is a fact cannot be a reason for this fact.

However, the bootstrapping objection does not show that intentions cannot be used as explanatory reasons. For instance, imagine that Sarah is asked why she is buying large amounts of ham, cheese and bread. She may reply that she is in charge of a birthday party on Saturday and that she intends to make sandwiches for the guests. If Sarah had no reasons to intend to make sandwiches, she would have no reasons to buy ham, cheese and

50 See, for similar arguments, Wallace (2001) and Raz (2005).

bread. In this sense, choices –this is, the formation of an intention – can be offered as reasons if those choices are based on reasons:

> [W]hen there is a reason to do what you have decided, the fact *that you have decided* to do it cannot exhaust that reason. *That you have decided* to do something can only be a reason for doing it when there is some further reason for doing what you have decided to do. [51]

Forming an judgement requires to deliberate about goals and means, to infer how to reach a state of things by performing a series of actions. It is not possible to deliberate without reasons. So, in order to form an intention based on practical reasoning, it is necessary to take a fact as a reason for that intention. Bratman[52] analyses this previous requirement as self-governing policies, which asses the agent on what to consider a reason for action. The Toxin Puzzle is a clear example of this constraint in the formation of intentions: there are coherence requirements between beliefs and intentions, which relate to the facts the agent takes as reasons (to act and to believe): "you cannot intend to act as you have no reason to act, at least when you have substantial reasons not to act"[53]. It is not possible then to intend to φ if the agent has no reasons to φ, at least concerning future-directed intentions.

By intending, some facts are used as reasons for performing certain actions; but these reasons are not created, as long as we do not create these facts. Intending to φ is a mental state, not a fact; to use the fact that we intend to φ as a reason, it is required to form that intention (for the fact to be actually what is the case), and forming that intention is the result of a deliberative process about the reasons for performing φ; so the agent does not actually create reasons for performing φ, but are recognizes some facts as reasons to φ[54]. Hence, the bootstrapping objection is a good argument against the claim that intentions create new reasons for action, but does not affect the claim that intentions, this is, the fact that we intend to φ, can play the role of reasons for action, including that same action φ.

---

51 Cullity (2008: 66).

52 Bratman (2004).

53 Kavka (1983: 35).

54 See van der Torre and Tan (1999); Kolodny (2005).

Let's state the difference between recognizing and creating reasons as follows. In the former case, the agent uses a fact to infer a conclusion by applying a valid inference rule. Some of these inferences may be practical inferences, and thus would justify another fact, an action to be performed. For example, let's suppose that Sarah watches the weather forecast on television, and it announces a rainy and windy weekend. A couple of hours later, an uninformed friend of Sarah proposes her to go for a picnic on Saturday; she might reasonably argue that going for a picnic is not a good idea, because horrible weather conditions are expected. Has Sarah created a reason for action? The fact that it will rain had not been used as a reason for action before her friend's proposal, so, in a way, she has turned a fact into a reason by using it in a practical reasoning process; similarly, it is possible to turn a fact into a premise through its introduction in an argumentative schema. However, Sarah has not *created* that reason. For Sarah to create a reason for action, she has to perform an action. Imagine that Sarah uses a pre-commitment strategy[55] to loose weight. She actively avoids buying junk food and ready-made snacks, and buys fruit and vegetables instead. The added cost of satisfying a craving for food is supposed to be a further reason for not eating unhealthy food –maybe not a normative reason, but a motivational one. This is precisely the point of using pre-commitment strategies.

In brief, an intention is a reason for action as far as the fact of intending can be used to explain or to justify an action; intending, as a mental state, cannot be a reason. However, it is important to note that, by intending, one does not create new reasons for action, this is, reasons that did not exist before forming the intention[56]. But it does not follow that the fact of intending does not have normative force.

55 I am using the concept of pre-commitment strategy as stated by Elster (2000); (2003).

56 Deciding is an action that can actually reassess the weight of reasons that the agent previously had for and against performing what is decided see Cullity (2008). But it is not entailed that, by deciding, an agent creates new reasons: if the agent has no reason to choose neither of her possible options, then by deciding to choose one option she cannot create a new reason.

2.2. Rationality requirements

Practical commitments are subject to normative constraints. By judging that she ought to φ, or intending to φ, an agent acquires "rational obligations", so to say. Amongst other things, she should find the available means and intend to perform what means she considers best; she should not revise her intention unless she has a decisive reason to do so; and she should perform the action intended, amongst others. Imagine an agent, Hannah, a philosophy graduate student, who is planning her academic year. She reads a call for papers through the *philos-l* mailing list, and decides that, all things considered, she should apply. She works on the topic, the venue is an interesting place, and she has funding opportunities. Also, she will have the opportunity to discuss her work with colleagues, and attending that particular conference would improve her CV. Nevertheless, as the deadline approaches, Hannah does not start writing her abstract proposal. She believes that, everything considered, she should apply, but she does not intend to apply: she has in fact decided not to apply. Here, Hannah would be considered akratic. Hannah's story does not end here; suppose that she has formed the intention to apply for the conference; she plans to read some recent papers on the topic and write the abstract. So far, she is being rational. However, the deadline is approaching, and Hannah has not started to write her abstract yet; she has failed to intend the means for achieving her intended end. Here, Hannah's irrationality is due to weakness of will, a violation of the means-end coherence requirement: a rational agent intends the means of her intended ends. Rationality, thus, imposes some restrictions to agency in order to be considered *rational* agency.

There are other constraints that rationality imposes to agency. Rationality requires a *rational balance* among beliefs, intentions and actions:

> An autonomous agent should act on its intentions, not in spite of them; adopt intentions it believes are feasible and forego those believed to be infeasible; keep (or commit to) intentions, but not forever; discharge those intentions believed to have been satisfied; alter intentions when relevant beliefs change; and adopt subsidiary intentions during plan

formation. [57]

It is widely accepted that rationality imposes such constraints. However, the way in which these constraints are expressed is controversial. I am interested in rationality constraints insofar they reflect the normative entailment between practical commitments and rational agency. This is, once Hannah has decided to submit an abstract to a conference, her decision to act normatively commits her to other things, such as engaging in deliberation about the means, intending (and doing) these means, or not forming other intentions that would make impossible to achieve her goal. Volitional commitment explains why Hannah actually *writes* an abstract and submit it, given her intention to apply for the conference. Normative commitment, on the other hand, explains why she *should write* the abstract –or revise her intention–, this is, why it is rational for her to do so. I will focus here on three constraints imposed by rationality through the adoption of normative commitments: enkrasia, resolve, and means-end coherence.

There are different ways in which these requirements can be understood: wide-scoped or narrow-scoped, using further normative concepts besides ought and reason, or through drawing a parallelism between practical and theoretical rationality. In what follows, I will present the two main views on the scope of rational requirements, and discuss some objections that have been raised against each of them. My suggestion is that narrow-scope has the ability to avoid two problems affecting wide-scope formulations: symmetry and infinite regress in choice. Second, I will argue that the enkratic requirement is valid if understood as a form of restriction. Requiring an agent to intend to do what she believes she ought to do is too demanding as a rationality conditions; enkrasia, I will argue, requires that the agent does not intentionally contravene her own judgements. Thus, the lack of intention does not violate this requirement. Lastly, I will propose that weakness of will constitutes a violation of the *resolve* requirement, which demands coherence and consistency amongst intentional states, specially between future and

57 Cohen and Levesque (1990: 214).

present-directed intentions. The *means-end coherence* requirement would then be derived from the resolve requirement.

2.2.1. Narrow-scope and wide-scope

The two traditional normative concepts are *reasons* and *ought*. In Section 2.1.3 I explained how the bootstrapping objection shows that, by intending, one does not create new reasons for acting as intended. However, it can be argued that it is irrational to intentionally contravene one's best judgement, as this would be a case of akrasia. This irrationality, Broome argues[58], cannot be considered a violation of an ought, because it is possible that the agent (mistakenly) judges that she ought to φ, but in fact she ought not to φ. Hence, Broome argues, there is a need for a further normative concept, which he calls *normative requirement*[59]. Rationality requires that judgements and intentions are consistent. He argues that rationality is subject to normative requirements; to act following such requirements is called *enkrasia*. It requires of the agent to intend to φ at time t if (1) she believes at t that she herself ought to φ, (2) she believes at t that, if she herself were then to intend to φ, because of that, she would φ, and (3) she believes at t that, if she herself were not then to intend to φ, because of that, she would not φ. Clauses (2) and (3), Broome says, mean that it is up to the agent whether or not to φ. The role of enkrasia is to make the results of deliberation practical, so beliefs about what the agent should do and her intentions to do it are connected[60].

Under one reading an enkratic agent does what she judges best; for now, I will use, in order to discuss narrow and wide-scoped formulation, this notion of enkrasia[61].

---

58 Broome (1999); (2007); (2010).

59 I am not going to defend here, nor criticise, Broome's arguments for the need of a third normative concept. The definition of a normative requirement is the following: *p requires q* is equivalent to *it ought be the case that* $(p \rightarrow q)$.

60 As long as I do not take the conclusion of practical reasoning to be an intention, I do not take enkrasia (nor akrasia) to be a normative property of practical reasoning, but a property of practical agency. Akratic agents reason correctly; they fail at another level, i.e. at implementing their judgements into intentions.

61 Under an narrower reading of enkrasia, which I will defend below, an enkratic agent does not intend to do something that contradicts her best judgement.

Rationality requirements can be formulated in a variety of ways, so an agent would comply with those requirements when doing what she judges she ought to do. A recent debate regarding the possible formulations of rational requirements concerns their logical structure. In broad terms, there are two opposing alternatives: wide-scope and narrow-scope formulations[62]. Both wide and narrow-scope formulations can be applied to the three kinds of rational requirements that I pointed out above: enkrasia, resolve, and means-end coherence. The distinction goes as follows:

> NARROW-SCOPE ENKRASIA: If you believe that you ought to φ, then you are rationally required to intend to φ.
>
> WIDE-SCOPE ENKRASIA: Rationality requires that [if you believe that you ought to φ, then you intend to φ].

Regarding the requirement to be resolute:

> NARROW-SCOPE RESOLVE: If you intend to φ, then you are rationally required to intentionally φ.
>
> WIDE-SCOPE RESOLVE: Rationality requires that [if you intend to φ, then you intentionally φ].

Applied to the requirement of means-end coherence:

> NARROW-SCOPE MEANS-END COHERENCE: If you intend to φ and believe that ψ is necessary for φ, then you are rationally required to intend to ψ.
>
> WIDE-SCOPE MEANS-END COHERENCE: Rationality requires that [if you intend to φ and believe that ψ is necessary for φ, then you intend to ψ].

The difference between narrow and wide-scope lies in the possibilities that an agent has available when she finds herself in a situation of irrationality. Regarding enkrasia, an agent

62 This debate has generated a growing amount of articles; the beginning of the discussion can be found in Broome Broome (1999); (2007), and Kolodny (2005) offers an interesting critique of Broome's defence of wide-scoped formulations—see Brunero (2010) for a recent critique of Kolodny's account. Way (2010) suggests a third approach, which would be medium-scoped, regarding the requirement of means-end coherence. A similar proposal was made by Setiya (2007a). They argue that the rationality requirement can be formulated as follows: if one believes that doing ψ is a necessary means to doing φ, then it is required that one [give up one's intention to do φ or adopt the intention to do ψ M]. Although this suggestion has some advantages over the wide-scope account, it still faces the asymmetry problem stated below.

can realize that she judges that she should do something, and that she has not formed the intention to do it. Following narrow-scope enkrasia, she is required to intend to do it. In our example above, Hannah, having judged that she ought to apply for the conference, but not intending to apply, is required to form that intention. This is: the only way Hannah can be rational is by forming an intention. However, from a wide-scope perspective, Hannah has a choice: she can either form the intention to apply for the conference, or revise (and, ultimately, abandon) her judgement that she ought to apply for the conference.

Now, regarding the resolve requirement and the means-end coherence requirement, the difference between narrow and wide-scope is similar: narrow-scope leaves the agent no choice, while wide-scope leaves open two possibilities for solving the irrationality problem. On the one hand, if an agent finds herself in a situation in which she intends to do something, but nonetheless she does not do it, she can either drop her intention, or do it. There is one more possible alternative in the case of the means-end coherence requirement. Following our example above, Hannah intends to apply for the conference but, nonetheless, she has not formed the intention to write the abstract. From a narrow-scoped view, Hannah ought to intend to write the abstract. A wide-scope formulation would let Hannah choose amongst three options: either she forms the intention to write the abstract, or she abandons her intention to apply for the conference, or, finally, she revises (and eventually abandons) her belief that writing the papers is necessary for applying for the conference.

Thus, applied to practical commitments, the question would be whether having such commitment makes it true that we ought to do what we are committed to, or that abandoning the commitment is also a permissible way to act rationally. In the former case, commitments would normatively affect other dependent beliefs and intentions, constraining and restricting their formation, reconsideration, or abandonment. In case of adopting a wide-scope formulation, the normativity of commitments would be hierarchically at the same level, so to say, as those other beliefs and intentions. Thus, if we are committed to do something, abandoning this commitment and committing ourselves

to the necessary means would be equally rational. I will now turn to the problem of each account, and argue for a narrow-scope formulation of rational requirements.

Narrow-scope formulations of the enkratic requirement are criticised for the counterintuitive consequences of cases in which the antecedent (the belief that one ought to do something) is true, but irrational: it can be a product of wishful thinking, or "obviously crazy"[63], it can refer to morally prohibited actions, such as intending to kill someone[64], or it can be a simply irrational belief about what one ought to do, such as spitting on Las Meninas[65]. In fact, most of the supporters of the wide-scope account reject narrow-scope formulations because of the possibility of being required to form the intention to do something irrational.

On the other hand, the main problem for wide-scope is closely related to its virtue: it is symmetrical. Wide-scope enkrasia rationality requires that you either intend to do something, or to abandon the belief that you ought to do it, all things considered. The symmetry problem is, precisely, that rationality does not seem to be symmetrical: is it really equally rational for an agent to drop her belief about what is best, and to form the intention to do what she judges to be best? Schroeder points out that wide-scope accepts as rational the possibility of rationalization, understood as changing one's belief in the light of the absence of intention:

> The first problem for Wide-Scoping is that it is symmetric. It doesn't distinguish between acting in accordance with your moral beliefs and adopting moral beliefs in accordance with your actions, and as a result it fails to distinguish between following your conscience and the distinctive vice of rationalization. Rationalization is the vice of changing your beliefs about what you ought to do, because you are not going to do it, anyway. According to the Wide-Scope view, this is precisely as good a way of satisfying this requirement as is actually paying attention to what you believe and acting accordingly. [66]

Imagine that Joe is considering quitting smoking. He judges that he should stop smoking but, when offered a cigarette after a few hours of experiencing withdrawal symptoms, he

63 Way (2011).
64 Hussain (2007).
65 Shpall (forthcoming), forthcoming.
66 Schroeder (2009: 227).

accepts it. To warrant the rationality of his behaviour, he abandons the belief that he ought to give up smoking. It seems that, in spite of complying with the wide-scope formulation of the enkratic requirement, Joe is not being completely rational. A resolution such as quitting smoking is a kind of intention whose role is precisely to block further reconsideration[67]. Joe is changing his belief that he ought to quit smoking because he has accepted a cigarette –which does not seem fully rational.

In the case of silly intentions, the asymmetry also holds. Suppose that Sam Shpall believes that he ought to spit on Las Meninas (for no particular reason). According to the wide-scope view, Shpall can cease to believe he ought to spit on Las Meninas. But once he comes to believe that he ought to spit on Las Meninas, without realising that he has no reason to do so[68], it seems that reconsidering his belief for no reason is just as irrational as forming it. It is also counterintuitive that he reconsiders his belief because he wants to comply with the wide-scoped enkratic requirement, because, was Shpall caring for rationality, he would wonder about his reasons to spit on Las Meninas in the first place.

Finally, there are cases in which an agent believes she ought to do something that, objectively, she ought not to do. Suppose that Bert believes that he ought to kill Ernie. Bert considers that he has reasons to do so: he has been blackmailed and threatened by his flatmate Ernie[69]. Tired of the suffering, Bert judges that the best thing he can do is killing Ernie. However, Bert has not formed the intention yet. From a wide-scope perspective, Bert can either intend to kill Ernie, or to drop his belief that he ought to kill Ernie. Narrow-scope critics seem to be somehow worried about the proposition "Bert is required to intend to kill Ernie" being true, so giving Bert the possibility of dropping his belief and still be a rational agent is a reassuring alternative. Given that Bert ought not to intend to kill Ernie, then he has to choose between two alternatives: remain irrational, or changing his belief about what he ought to do. Bert wants to be rational; then, he is required to

67 Holton (2009). See also Chapter 1.

68 Let's assume that this is possible.

69 An objectivist would argue that Bert has no reasons to kill Ernie, this is, that being blackmailed and threatened by Ernie is not a reason to kill Ernie. But, as I argued above, this would be an evaluation of Bert's rules of practical reasoning and evaluation mechanisms, which determine what to treat as a reason. Thus, for now, I will assume that Bert has (good or bad) reasons to kill Ernie.

change his belief. Thus, wide-scopers would in fact be arguing that there is no choice for Bert: he ought to change his belief. On the other hand, it seems plausible to assert that one ought not to change her beliefs without having any reason to do so: having new evidence, realising a mistake in reasoning, and such. In our example, Bert is required to change his belief on the sole basis of the wrongness of the content; but, if Bert does not have access to reasons for considering this action wrong, then his change of belief would not be made on a rational basis.

Besides the symmetry problem, there is another reason why wide-scope formulations are controversial. They state that complying with the requirements of rationality is a matter of choice. An agent, facing an akratic state, can either change his belief that she ought to φ, or form the intention to φ. Or, she can remain in an irrational state[70]. Letting aside this last possibility, the agent is then facing a choice. The agent, in order to make a rational choice, has to evaluate what reasons for choosing any of the alternative she has. Imagine that she judges that she ought to ψ [change her belief that she ought to φ]; is she required to intend to ψ? No: she can choose between intending to ψ or abandoning her belief that she ought to ψ. Choosing without an anchor leads to two possible scenarios: infinite regress, or arbitrary choice. The agent needs a point from which to deliberate and choose. Usually, this point is the agent's initial aim, goal or intention: this is why critics of wide-scope formulation argue that it fails to account for the *directionality* of deliberation and choice.

Given these two problems of wide-scope accounts, I believe narrow-scope formulations provide a better account of the rationality requirements stated above. The only objection against narrow-scope is that it allows for asserting that an agent ought (in a subjective sense) do something that he ought not (in an objective sense) to do[71]; he ought

70 Kolodny (2005) has put forward the question of whether there are reasons to comply with rationality requirements; although his approach is very interesting, I will not discuss it here for it would exceed the scope of this work.

71 Schroeder (2009).

(objectively) to revise her judgement instead. Schroeder suggests that the relationship between subjective and objective oughts can be defined as follows:

> **[S]ubjective ought test :** X subjectively ought to do A just in case X has some beliefs which have the following property: the truth of their contents is the kind of thing to make it the case that X objectively ought to do A. [72]

The fact that agent objectively ought to do something is taken to be an agent-neutral claim, that can be either truth or false depending on how the facts stand. To put it otherwise: if fact X is an objective reason to φ, then anyone ought to φ if fact X is the case. On the other hand, if an agent believes that the fact X is the case (although it is not), that particular agent subjectively ought to φ. Objective oughts have nothing to do with the agent's beliefs: an agent ought to drive on the right side of the road independently of what she believes.

> [W]hen we know all of the relevant facts, what we ought rationally to do is the same as what we ought to do in the decisive-reason-implying sense. But when we are ignorant or have false beliefs, these oughts may conflict. Suppose that, while walking in some desert, you have disturbed and angered a poisonous snake. You believe that, to save your life, you must run away. In fact you must stand still, since this snake will attack only moving targets. Given your false belief, it would be irrational for you to stand still. You ought rationally to run away. But that is not what you ought to do in the decisive-reason-implying sense. You have no reason to run away, and a decisive reason not to run away. You ought to stand still, since that is your only way to save your life. Some people would say that you do have a reason to run away, which is provided by your false belief that this act would save your life. But if we say that false beliefs can give people reasons, we would need to add that these reasons do not have any normative force, in the sense that they do not count in favour of any act. And we would have to ignore such reasons when we are trying to decide what someone has most reason to do. It is better to describe such cases in a different way. When we have beliefs whose truth would give us a reason to act in some way, we have what I shall call an apparent reason to act in this way. If these beliefs are true, this apparent reason is also a real reason. If these beliefs are false, we have what merely appears to be a reason. In the case of the angry snake, given your false belief that running away would save your life, you have a merely apparent reason to run away […] But what it would be rational for people to do depends on their apparent reasons, whether or not these reasons are real, or merely apparent. [73]

Parfit identifies the distinction between objective and subjective oughts with the difference between reasons and rationality. Rationality is agent-relative (or perspective-dependent): it depends on the agent's beliefs. It would be irrational for an agent not to act upon her

72 Ibid., 230.

73 Parfit (2011: 1:34–5).

reasons, whether they correspond to facts or not, this is, whether they are "apparent" reasons. However, Parfit argues, reasons are perspective-independent: they stand in a normative relation with certain actions, independently of whether they are believed to be true by any agent.

It is quite counterintuitive to assert that, while reasons are perspective-independent, rationality is perspective-dependent[74]. This double perspective regarding reasons and rationality can be stated as follows. On the one hand, if an agent has reasons for doing something, it is rational for her to intend to do it, regardless of the quality of her reasons –rationality is therefore perspective-dependent. On the other, only true reasons (objective reasons, this is, normative reasons that refer to facts, and that it is objectively correct to use as reasons in practical reasoning to infer that particular conclusion) count as reasons: if an agent, Hannah, believes that her glass is full of gin and tonic, and it happens to be gasoline, then Hannah has no reason to drink from that glass –reasons are therefore perspective-independent. Assuming that both are perspective-independent leads to a strange situation: it is rational for Bert not to kill Ernie (although he thinks he should), and it is rational for Hannah not to drink from the glass (even if she thinks it has gin and tonic). But assuming that rationality is perspective-dependent and reasons are perspective-independent (which is labelled by Gibbons "The Bad") opens a breach between the normativity of reasons and the normativity of rationality: "Since the defining feature of The Bad is the gap between normative reasons and rationality – the former are perspective-independent while the latter is perspective-dependent – whatever positive normative status is conferred by normative reasons, that status cannot be rationality"[75].

Let's bring back Bert's example. He has reasons to murder Ernie. We, or an external observer, might consider Bert's reasons as good or bad reasons. Arguing that Bert's reasons are not reasons at all does not say anything about reasons, but about the rules with which an agent is entitled to infer she ought to do something. This is: objective

74 See a full blown defence of this claim in Gibbons (2010).

75 Ibid., 345

reasons are not reasons, they are inference norms, which express rules that are social, moral, instrumental, etc. (see §2.1.1). Thus, I am not sure about why stating that Bert ought to kill Ernie necessarily refers to Bert having objective reasons (this is, subjective reasons based on objective rules) to kill Ernie. The only thing we can infer from Bert having reasons is that, if they are reasons for Bert, then Bert ought to judge that he ought to kill Ernie. If oughts and requirements are perspective-dependent, which I believe to be the case, then creating a new normative concept seems unnecessary. On the other hand, the concept of normative requirement is attractive because it gathers that rationality is requiring something from you in order to be rational; he concept of ought expresses the requirement, but it does not make clear who or what is requiring things from you. For example, you are required to comply with social norms, but rationality has nothing to do (broadly speaking) with the content of those norms, nor is requiring you to comply with them.

To sum up, the symmetry of the wide-scope formulation fails to consider the rational constraints to belief change. Wide-scope accounts try to avoid the possibility of being required to do something that, objectively, one ought not to do in the first place. Wide-scope emerges as an objectivist solution to the problem of coherence between beliefs and intentions. If an agent ought not to believe that she ought to φ, then she ought not to φ: she ought to change her belief instead. However, the problem of the conditions under which it is rational to change one's mind is eluded. On the other hand, the problem of detaching the requirement given the antecedent is not problematic if normative requirements, as well as oughts (which I take to be the same thing), are understood as perspective-dependent.

2.2.2. Normative requirements and practical commitment

My aim now is to argue that the requirements of practical agency are better understood as *directional* and *narrow-scoped*. I will first explore how the enkratic requirement ought to be understood, in order to correctly reflect why its violation entails akratic irrationality.

My suggestion is that enkrasia is better formulated as a restriction (forbidding the formation of a particular intention) rather than a requirement (demanding to form a particular intention). Second, I will argue that weakness of will is irrational insofar it is a violation of the resolve requirement (of which the means-end requirement is derived). Weakness of will consists in a failure to act according to one's intention. This definition does not only cover cases in which the agent holds conflicting intentional states, but also cases in which she fails to form the subsequent present-directed intention.

Akrasia and enkrasia

Akrasia (sometimes referred in the literature as weakness of will; see §1.2.3 for an argument for making the distinction) is sometimes described as a failure to comply with this formulation of enkrasia[76]. Davidson[77] followed this tradition, and defended that weakness of will consists in the failure to move from an all-things-considered better judgement to an intention to do what is judged to be best. Traditional conceptions on akrasia take it as a violation of a commitment of the agent: "[for traditional conceptions] an agent who decisively judges it best to A is thereby rationally committed to A-ing, in the sense that (as long as the judgment is retained) the uncompelled, intentional performance of any action that he believes to be incompatible with his A-ing would open him to the charge of irrationality"[78]. This tradition has continued until nowadays; for instance, McIntyre claims that "familiar examples of weakness of will display three notable defects: (1) the failure to do what one has judged it best to do, (2) the failure to do what one has most reason to do, and (3) the incoherent attitudes of agents who fail to do what they believe they ought to do"[79]. Similarly, Wedgwood defines akrasia as "willingly failing to do

76 The problem of akrasia goes back to Plato, who argued for the impossibility of acting against one's best judgement (Protagoras, 358 b-c). For an historical overview of weakness of the will and akrasia, see Gosling (1990) and Thero (2006: 183:).

77 Davidson (1980).

78 Mele (1995: 71).

79 McIntyre (2006: 290).

something that one judges that one ought to do"[80]. This claim is known as *normative judgement internalism* (NJI):

> ENKRASIA (NJI): Necessarily, if one is rational, then, if one judges 'I ought to φ', one also intends to φ.

NJI claims that there is an internal link between an agent's normative judgements and her disposition to act. In fact, those who defend that the conclusion of practical reasoning is an intention are committed to some version of NJI, because it provides the justification of the link between a normative judgement and the formation of an intention. For example, Stroud argues that, "through deliberating [...], an agent can reach a conclusion –a judgement– which has an internal, necessary relation to subsequent action or intention"[81]. I have argued in §2.1.2 that the conclusion of practical reasoning is a normative belief. I will argue in the next Section that rationality requires an agent to intend accordingly only when she already holds a previous intentional state. Therefore, normative beliefs are not internally linked to the formation of an intention, although intention formation is rationally limited by the normative judgements the agent accepts.

A second formulation of akrasia defines it as acting against one's better judgement. In fact, most accounts of weakness of will or akrasia agree with this claim –the difference with the former formulation is that not every author that agrees with this view also agrees with NJI, while most proponents of NJI agree with this formulation. From this perspective, akrasia entails acting intentionally while, at the same time, judging that one ought not to do what in fact is being done[82]. With regard to the mechanisms that make akrasia possible, Mele argues that "the motivational force of a want may be out of line with the agent's evaluation of the object of that want"[83]. This is, the assessment of the reasons that would justify an action and our desires of performing it need not to coincide. In fact, following Mele, motivational attitudes towards one's judgements are necessary in order to produce a corresponding intentional state:

---

80 Wedgwood (2007b: 25).

81 Stroud (2003: 122).

82 See Mele (1987); Audi (1993); Gilead (1999); Tenenbaum (2010).

83 Mele (1987: 37).

> [A]ttributing an action-guiding function to evaluative judgements […] does not commit one to supposing that the judgements are themselves *logically* or *causally* sufficient for the presence of corresponding intentions. […] There is no motivational magic in the thought *content* "My A-ing would be best." [84]

Similarly, Audi points out that a judgement may lack motivational force:

> Practical reasoning, then, is a process by which agents infer judgments favoring action from premises expressing motivation and instrumental cognition. Normally, they have sufficient motivation of the kind in question—whether its basis is self-interest, duty, emotion, or something else again—to enable their concluding judgment to produce action or intention. But that judgment can provide normative guidance for conduct even where, as with weakness of will, it lacks sufficient motivational force to yield action. [85]

Failing to intend to do what one judges best is different from acting intentionally against one's better judgement. It is far from clear that, by not intending to do what she believes she ought to do, and agent holds inconsistent beliefs and intentions. In fact, the distinction between the absence of intention and the presence of a negative intention is often overlooked. This is, *intending not to φ*[86] is quite different from *not intending to φ*. I assume that the difference is clear and unproblematic; my claim is that this difference is relevant for the understanding of the normative relation between judgements and intentions. In fact, the problem with NJI is that it assumes that it is irrational *to believe that I ought to φ* and, at the same time, *not to intend to φ*. But I believe this condition is far too demanding.

Although, narrowly speaking, akrasia consists in *acting* intentionally against one's best judgement, holding a prior future-directed intention to act against one's best judgement also seems to violate the enkratic requirement. After all, the difference between holding a prior future-directed intention and a present-directed intention while the action is being performed is a matter of time. Of course, not every present-directed intention is preceded by a future-directed one. As I explained in §1.2, not every intentional action is

84 Mele (1995: 25); his italics.

85 Audi (2006: 81).

86 I assume here that intending not to φ, as well as not φ-ing intentionally, are proper intentional states, and that omissions are actions (although they might be merely mental actions, given that they are not exercised towards the world). See Clarke (2010) for a discussion on intentional omissions, and how they differ from absence of intention.

the result of an act of choice. Akrasia would also cover these situations. If I judge that I ought not to eat more than three chocolates a day, and when I open the box and eat three chocolates I go for the fourth without prior deliberation, I am being akratic (if my judgement has not changed)[87]. Therefore, although I do not agree in that an intentional state necessarily follows from a practical judgement (judgement internalism), I will defend a version of volitional internalism, which claims that an agent's choices and intentions do bear a internal relation to her practical judgements[88].

Akrasia, then, requires having formed a judgement, either positive or negative, about what one ought to do. Its irrationality lies in an incorrect relation with an intentional state of the agent. Given that an agent can intend to do something, or to intend to refrain from doing something (i.e. not to do something), and that she can also lack the relevant intentional state, the following combinations may obtain:

Bφ	Iφ	Enkrasia
	I¬φ	Akrasia
	¬Iφ	??
	¬I¬φ	??

B¬φ	Iφ	Akrasia
	I¬φ	Enkrasia
	¬Iφ	??
	¬I¬φ	??

Table 1: Combinations of judgements and intentions

It should be noted that I am using here the concept of intention in a broad sense, to refer both to actions that are made intentionally, and to future-directed intentions. Of course, there are differences between the two of them. Mainly, in order to avoid akrasia, an agent who holds a future-directed intention to do something she judges she ought not to do can drop her intention; an agent who is acting intentionally against her better judgement can stop acting.

87 In fact, many of the examples of akratic actions found in the literature refer to situations in which an agent succumbs to temptation, and does something intentionally that contravenes her practical judgements. In theses cases, it can be hardly said that the agent had a previous future-directed intention; rather, she does something intentionally but impulsively, this is, without previous deliberation.

88 The difference between judgement and volitional internalism is presented by Hinchman (2009).

There are four uncontroversial cases: two of akrasia, and two of enkrasia. Hannah believes that she ought to buy a new bicycle, and she intends to do so; similarly, Hannah believes that she ought not to spend money on a bicycle, and she intends to refrain from buying anything. In these two examples, Hannah displays enkrasia: her intentions and her beliefs are lined up. On the contrary, if Hannah believes that she ought to buy a bike, and she intends not to do so, she displays akrasia. The same goes for believing that she ought not to spend money, and she intends to spend it. However, what about the cases in which Hannah does not intend anything?

First of all, it is unclear in what sense not intending to φ and not intending not to φ are distinct. If Hannah lacks an intentional state, it does not matter much what the content of that state is, for it is something that Hannah lacks. Usually, when an agent intends to φ, she lacks the intention not to φ; and vice versa, if she intends not to φ, she lacks the intention to φ. But it may well be the case that she intends neither to φ nor not to φ. For example, I do not intend neither to go to Thailand next year, nor I do not intend not to go. It is simply an intention I do not have. Therefore, it is not true that every time an agent does not intend to φ, she intends not to φ, or that if she does not intend not to φ, then she intends to φ.

Following our example above, suppose that Hannah judges she ought to buy a new bike, all things considered. She does not intend to buy it; nor she does not intend not to buy it. She has suspended choice for whatever reason. I do not believe that she would be acting akratically: she is not acting against her judgement, nor holds an intention to do something against her judgement. This claim is clearer when it comes to judgements not to do something. Imagine, for instance, that Hannah believes that she ought not to smoke in a bar, because it is forbidden. In fact, Hannah does not smoke either: she has been asked by a friend whether she ought to smoke in the bar (if she wanted to), and she judges that, everything considered, she ought not to smoke there. However, Hannah does not need to form the intention not to smoke in the bar. She does not have the intention of

smoking there, and this is enough to act rationally. The lack of any intentional state regarding smoking does not contradict her judgement: she is not being akratic.

Formulating the enkratic requirement in a way that agents who judge that they ought not to do something are required to form an intention not to do it seems way too demanding. In many occasions, we do not deliberate about what we should do, but about whether we should do something in particular; and the answer can be negative, as well as positive. If I wonder now whether I should do anything unreasonable, for example, and I conclude I ought not, requiring to form the corresponding intention not to do it seems too strong as a condition for rationality. For example, suppose that I visit the Prado Museum, and standing in front of Las Meninas, I recall Shpall's article, and wonder myself whether there are reasons to spit on Las Meninas. After evaluating the reasons I have for and against it, I judge that I have no reason to do it, and many (normative) reasons not to do it. Therefore, I decide that I ought not to spit on Las Meninas. Am I required to form the intention not to spit on Las Meninas? Merely by not intending to do what I believe I ought not to do seems enough to avoid irrationality.

Hence, my suggestion is to formulate enkrasia as a restriction, rather than a positive requirement to form an intention:

> ENKRASIA (NARROW+): If you believe that you ought to φ, then rationality requires that you do not [intend not to φ].
>
> ENKRASIA (NARROW -): If you believe that you ought not to φ, then rationality requires that you do not [intend to φ].

This narrow-scope formulation states that the only thing that is required from an agent who judges she ought (not) to do something is that she does not intend to do something that violates her judgement[89]. To put it differently: you ought not to intend to do something that you judge you ought not to do. This is why I have some reservations about

89 In fact, this formulation is logically equivalent to a wide-scoped one relating the judgement and the intentional state (or its absence) through conjunction rather than implication, given that $\neg(p \wedge \neg q)$ is equivalent to $p \rightarrow \neg\neg q$. I prefer the narrow-scope formulation because it gathers the directionality of the normative constraints, but I find this version of the wide-scope also acceptable. The wide-scope formulation could go as follows:
ENKRASIA (WIDE -): Rationality requires that you do not [believe that you ought to φ, and intend not to φ].

considering enkrasia as a form of requirement: it is rather a prohibition. We are not required to do something, but to abstain from doing it. Of course, unintentional and non-reflective abstentions also count as complying with enkrasia. If I believe I ought not to lie, I do not need to form the intention not to lie every time I talk; not forming the intention to lie suffices for not violating enkrasia.

Furthermore, a narrow-scope formulation of enkrasia does not exclude the possibility of exiting from the requirement by reconsidering one's judgements. Given that narrow-scope formulations are conditionals, denying the antecedent makes the consequent no longer required. But, as Lord[90] points out, exiting from the requirement—i.e. making it no longer apply to us through denying the antecedent—is not a form of complying with it, but it does not violate it either.

This table represents the possible relations between judgements and intentional states, regarding enkrasia and akrasia:

Bφ		B¬φ	
Iφ	Enkrasia-derived	Iφ	Akrasia
I¬φ	Akrasia	I¬φ	Enkrasia-derived
¬Iφ	Enkrasia-compatible	¬Iφ	Enkrasia
¬I¬φ	Enkrasia	¬I¬φ	Enkrasia-compatible

Table 2: Akrasia and three kinds of enkrasia.

I have made a threefold distinction regarding enkrasia. In the formulation I have suggested, an agent is enkratic as long as she does not intend to do something she believes she ought not to do, and vice versa. Also, if an agent intends to φ, then she does not intend not to φ; otherwise, she would have inconsistent intentions[91]. Therefore, if an agent intends to φ, and believes she ought to φ, she is being enkratic, given that, by intending to φ, she is also not intending not to φ. The label 'enkratic-derived', thus, aims

90 Lord (2011)

91 Of course, it is possible for an agent to intend to φ and to intend not to φ at the same time; it is irrational, but possible. I am assuming here that the agent is rational.

to stress that the rationality of acting according to one's judgements is derived from a requirement of not holding inconsistent attitudes. Finally, the 'enkrasia-compatible' case is such that the agent believes she ought to φ, and she does not intend to φ, or vice versa. This case is a bit more complicated. Suppose that Hannah judges that she ought to φ, and at the same time she does not intend to φ. This absence of intentional state is both compatible with not intending anything at all (which does not violate the enkratic requirement) and also with intending not not to φ –which does violate enkrasia. Thus, merely by not intending to do what one judges best, an agent is not violating the enkratic requirement, but akrasia is not ruled out. We lack of sufficient information to know whether, having judged she ought to φ, the agent is being akratic (if she also intends not to φ) or enkratic (if she does not intend anything at all).

To illustrate this, let's bring back Shpall's example. Sam Shpall claims to judge that he ought to spit on Las Meninas for no particular reason: he has formed a "ridiculous" belief[92]. Shpall offers an interesting analysis of the normative bond between his belief and the intention he would be required to form in order to be rational. He argues that, besides reasons, oughts and normative requirements, there is a need for a further normative concept: rational commitment. He argues that he cannot be required to intend to do something irrational, but he can be committed to do so on the basis of his judgement. This possibility allows for the differentiation between internal and external rationality constraints. Shpall claims that, by judging, one is internally committed to intend, but rationality requirements refer to the external (objective) bond, and thus they also require that the judgement is rational in the first place. I have two objections to this view.

The first objection is that I take all rationality constraints to be internal, as I argued above. Basically, I believe that rationality constraints are formal—objectivist views defend that they are not only formal, but also material. It does not matter how coherent or consistent an agent's reasoning is: she is, additionally, rationally required to grasp the

92 This was in fact a possibility that I have not taken into account in my defence of the claim that the conclusion of practical reasoning is a normative belief. Although I have serious doubts about whether an agent can form this kind of ridiculous beliefs just because, I will assume that it is possible to do so.

difference between objectively good and bad reasons. Although this difference is indeed relevant for ethical theory, I do not think that a theory of the normative force of judgements and intentions ought to include it. Thus, I do not find problematic to state that someone is rationally required (ought) to do something stupid, or unethical. I do not take this claim to express any objective obligation the agent is subject to, and so claiming that "Shpall ought to spit on las Meninas" (if this were correctly inferred from Shpall's reasons and intentions) is not defining an objective obligation of Shpall. Anyway, below I will argue that Bert ought to kill Ernie—i.e. that he is rationally required to do so, so I will further clarify my argument when analysing that example.

The second objection is that, even if all rationality requirements are internal, or some of them external and others internal, Shpall is not required, neither internally nor externally, to intend to spit on Las Meninas. As I have argued above, the enkratic requirement does not demand to form an intention to do what one judges she ought to do, but forbids the agent to form the intention not to do it. Regardless of the means by which Shpall has come to believe that he ought to spit on Las Meninas, his belief is not connected with any previous commitment Shpall had (as far as Shpall says). I take Shpall's example to be of the same kind as exploratory deliberation, at best. Shpall would be normatively required to spit on Las Meninas if his belief depended on a previous commitment of his. For example, suppose that Shpall is committed to destroy Velazquez's artwork. He believes that it is a great mistake to admire a painter who Shpall takes to be mediocre and overrated. He has built a plan: to visit all the museums in which Velazquez's paintings are exhibited, and mess with them. Stabbing a baroque masterpiece is too risky: Shpall does not want to expose himself to being arrested. Spitting seems a safe alternative. He visits the Prado museum, and he believes he ought to spit on Las Meninas. In this case, Shpall would be rationally required to do so, because his belief serves a prior commitment; not doing so would not be a violation of the enkratic requirement, but a violation of the means-end coherence requirement. To sum up: Shpall is not affected by

enkrasia not because his belief is irrational[93], but because his judgements is not related to any intention he had.

Weakness of will, resolve and means-end coherence

While akrasia is the failure to hold consistent judgements and intentions, weakness of will refers to the failure to hold consistent intentional states[94]. It is a rational failure of self-control and willpower: the agent's future-directed intentions do not properly exert control over further intention formation. An agent can display weakness of will both because of her present and her future-directed intentions. Following our example above, suppose that Hannah intends to apply for a conference. Then, a friend of her invites her to spend a week at her house, which is a holiday residence by the sea. Hannah knows that, if she accepts the invitation, she will not write the abstract. Despite her previous intention to apply for the conference, she starts packing. She has not dropped her previous intention, but formed a further future-directed intention to go to the beach for a week. There are different ways by which Hannah can rationalise her having those two incompatible intentional states. She can wishfully think that she will write the abstract in the beach, or that she will be able to write the abstract the night she arrives home, so she can send it before the deadline. Or she can reconsider her previous intention (applying for the conference) without reconsidering the broader goal that made her intend to apply for the conference in the first place, i.e. the intention to improve her CV. Thus, she might experience remorse, or regret, for abandoning the intention to write the abstract.

On the other hand, an agent can have conflicting future and present-directed intentions. For example, Hannah want to lose some weight, and thus she intends to go on a strict diet. One day, she is having lunch with a friend, and stands in front of her friend's

93 Way (2009) suggests that the rationality of the normative belief "I ought to φ" is necessary for the enkratic requirement to apply; holding irrational beliefs, as it would be the case of Shpall's belief that he ought to spit on Las Meninas, does not require that the agent forms the intention to do what he (irrationally) believes to be best. Contrary to Way, my point is that the rationality of that belief is irrelevant, for what matters is its connection to a previous commitment.

94 See §1.2.3; also Dodd (2009).

chocolate cake. Without much deliberation (or without any at all), Hannah takes the spoon and eats a bit of chocolate cake. In this case, Hannah is intentionally doing something (eating cake) that contravenes a previous future-directed intention of her. Again, there are different ways through which Hannah can rationalise her action. She may temporally drop her future-directed intention, this is, suspend her diet for that lunch. Or, she can think that one spoon of cake does not affect her intention to diet.

As I explained in §1.2.3, weakness of will is caused by failures of self-control, usually accompanied by a shift in the motivations of the agent. The difference between akrasia and weakness of will lies in that our practical judgements do not require that we intend to do what we consider we ought to do; on the contrary, our intention to φ does require that we intentionally φ. This also entails that we are required to do (and to intend to do) what we consider we ought to do in order to achieve φ. Otherwise, we would be intentionally doing not φ. Having an intention requires a certain response from the agent, specifically, to form the relevant implementation intentions[95]. An agent displays weakness of will, Pettit[96] argues, when she holds by intentional states in the light of which a certain response (an action) is required, and nonetheless she fails to act in the required manner. Although means-end coherence has drawn much more attention in the literature, there is a more basic requirement that links intentions and actions, which I will call *resolve*[97].

> RESOLVE (NARROW+): If you intend to φ, then you are rationally required to intentionally φ.
>
> RESOLVE (WIDE-): Rationality requires that you do not [intend to φ and intentionally do not φ]

95 See Gollwitzer (1999).

96 Pettit (2003a). It should be noted that Pettit uses the concept of 'akrasia' to refer to this failure; I believe that the kind of irrational action he refers to is better characterized as weakness of will.

97 I borrow this name from Hinchman, who examines what he calls *resolve internalism*, "the thesis that intending, resolving or otherwise willing to φ bears an internal relation to actually φing (or at least attempting to)" Hinchman (2009: 396). Similarly, Cohen and Handfield use *resolve* to label the "capacity for resolute maintenance of one's intentions " Daniel Cohen and Toby Handfield (2010: 907). Although I will not use the term exactly in the same sense that these authors, I believe it is useful for expressing an internal relation between intending and doing.

The resolve requirement can be understood either as a wide or a narrow-scoped formulation. The narrow-scope formulation gathers the directionality of intentions: once the agent intends to do something, she creates a normative bound to her future actions. This is normative commitment implicit in practical commitments. However, the wide-scope formulation is logically equivalent. It claims that there are two states that a rational agent cannot hold simultaneously: intending to φ and intentionally not φ-ing. This requirement can be subject to further development, and there are several derived requirements. For example, if an agent intends to φ, and knows that by ψ-ing she will not be able to φ, then she is required not to intend to ψ. The means-end coherence requirement is thus a variation of the resolve requirement:

> MEANS-END COHERENCE (NARROW+): If you intend to φ and believe that ψ is necessary for φ, then you are rationally required to intend to ψ.
>
> MEANS-END COHERENCE (WIDE-): Rationality requires that you do not [intend to φ, believe that ψ is necessary for φ, and do not intend to ψ]

This requirement states that the agent's means and ends should be coherent: it is required that an agent intends the means she believes to be necessary for achieving her ends. As we have seen, this requirement cannot be based on the capacity of intentions to create new reasons, for that would be a case of bootstrapping. This is, that an agent should intend the means not because intending the end is a reason to do so (a reason that did not exist before intending it), but because it would be a violation of the resolve rational requirement.

Therefore, while enkrasia is compatible with the lack of intention, resolve requires that the agent holds a present-directed intention to fulfil her future-directed intentions. A violation of this requirement can take several forms; for instance, by not forming the present-directed intention—in which case an absence of intention also counts as a violation—or through the formation of a future-directed intention that prevents the achievement of a prior (and not abandoned) future-directed intention. Weakness of will is a failure that concerns the consistency and coherence amongst intentional states, on the one hand, and the controlling function of intentions, on the other. While the absence of

intention does not contradict a practical judgement, the absence of intention entails that the agent is intentionally doing something (whatever she is doing) that prevents her from fulfilling her prior future-directed intention. As I have argued in §1.2, practical commitments, of which future-directed intentions consist, entail a volitional commitment, this is, a bond between the agent's intentions and actions through the control exerted by intentional states. If the agent fails to exert such control without abandoning her intention, she can be branded irrational. This claim holds in cases in which the agent has not suffered a total loss of control, because in these cases, the alternative action that prevents the agent from achieving her prior goal can hardly be described as intentional. Thus, practical commitment does not only imply that an agent's future-directed intentions exert control over her further intentional states and behaviour, but also that they ought to do so.

Although the wide-scope view also appropriately expresses the logic inconsistency amongst intentions, the narrow-scope formulation gathers a feature of the normativity of intentions that the wide-scope formulation does not: the directionality of the normative and volitional power of practical commitments. By intending to φ, an agent is normatively committed to attempt to φ, this is, to hold the present-directed intention to φ[98]. The incorrectness of being committed to φ and not doing φ cannot be equally solved either by ceasing to intend to φ or by doing φ. The hierarchical structure of intentions explains why being committed to φ imposes some normative constraints on the agent, which of course cease to exist if the agent ceases to be committed. Complying with a

98 Lorini and Castelfranchi (2004) argue that, while attempting to do something entails that the agent intends to do that something, trying does not see also Bratman (1987). On the other hand, Adams and Mele claim that "tryings are effects of the normal functioning of appropriate intentions" Adams and Mele (1992: 326). On their view, proximal intentions (this is, present-directed intentions) initiate tryings. Tuomela argues for a similar resolve requirement (although he prefers a wide-scope formulation) when analysing the group's commitments Tuomela (2007: 32); see also Tuomela (2000). He claims that intending requires trying. For an account of the trying / intending debate between Bratman and Tuomela, see Mele (2003b). I will consider here that the resolve requirement demands that future-directed intentions appropriately exert control over further present-directed intentions, tryings (in Adams's and Mele's sense) or attempts (in Lorini and Castelfranchi terms). However, I do not develop here a theory about the appropriate control exerted by future-directed intentions, but focus instead on conflicting intentional states, for each state prevents the achievement of the other. I am aware that my approach is limited, but entering into this debate would exceed the scope of this work.

requirement and making it not to apply to us (because we have ceased to be committed) are different tasks, which wide-scope formulations fail to distinguish. In a recent contribution, Lord[99] argues that exiting from a requirement is different from complying with it. A narrow-scope formulation is a conditional relation. If we make the antecedent false, then the relation does no longer apply to us, but we are neither complying with nor violating it. For example, Bert judges that he ought to kill Ernie (moved by reasons and prior commitments). Hence, he is required to intend to kill Ernie. However, if Bert ceases to believe that he ought to kill Ernie, he is no longer required to intend to do so: he has 'exited' from the requirement. One of the ways of exiting a requirement is through reconsideration of one's reasons. Sometimes, it is reasonable to do so; we may have new information, or our motivations may have changed. However, it is also possible that the agent changes her mind temporally, because of a preference reversal[100].

In many cases, an agent displays both weakness of will and akrasia; but it is possible to display only one of these rationality failures. Following the example above, Hannah can form the intention to go to her friend's house, knowing that it is incompatible with her previous intention to apply for the conference. At the same time, Hannah can judge that she ought to apply for the conference, given that she intends to improve her CV. In this case, Hannah would be displaying akrasia—because she intends to do something that contravenes her judgement—and weakness of will—because she holds conflicting intentional states. When the agent's deliberation is prompted and guided by previous intentions, an agent is supposed to intend to do what she judges best. In fact, some authors argue that practical reasoning is practical in virtue of serving to these prior intentions[101]. I have argued in §2.1.2 that an agent can voluntarily engage in practical reasoning, independently of whether that reasoning is made with the aim of choosing, thus forming an intentions, or it is just a case of exploratory reasoning. An agent can then be weak-willed without being akratic, if she does neither judge that she ought to do

99 Lord (2011).
100 Elster (2006); see also §1.2.3.
101 See Alvarez (2010b).

something, nor that she ought not to do it. For example, Hannah can form the intention of going on a diet without judging that it is the best thing she could do. She has also evaluated doing exercise and visiting a doctor, but she has not made up her mind. However, given that she really wants to lose weight, she chooses to go on a diet, even if she has not reached a conclusion about what she ought to do. Then, while having dinner with a friend, she takes the spoon and takes a piece of cake. Hannah holds conflicting intentional states, but she is not being akratic. Furthermore, it is possible for an agent to be weak-willed and akratic, but that the intention that contradicts a prior judgements is the first intention, not the second. Suppose that Hannah believes that, all things considered, she ought not to go on a diet. She is not fat, and so she does not need to self-impose such a restriction. However, contrary to that judgement, she chooses to go on a diet, because she wants to be thinner (than she believes she ought to be). Then, when facing a chocolate cake, she intentionally eats a bit, skipping her diet. Here, Hannah is being weak-willed, but also akratic, although not in a traditional manner, because it is her intention to go on a diet what enters into conflict with her best judgement.

I conclude this Chapter with a reflection on the relation between akrasia and weakness of will. I have tried to stress that, insofar the conclusion of practical reasoning is a normative judgement, it is not rationally required that the agent forms a subsequent intention after deliberation, even if she has reached a conclusion, and her reasoning is complete. I have focused on cases in which the agent is motivated to engage in practical reasoning by other than serving to an intention—exploratory reasoning, advice, or hypothetical scenarios. Also, I have pointed out that an agent can deliberate, while having reasons not to make a choice. However, it should be noted that, often, what prompts and guides practical reasoning is an active goal that the agent has, this is, an intentional state. In these cases, practical reasoning serves to find out the means through which to achieve the goal. Failing to form an intention based on judgements about what one ought to do in order to achieve the intended goals constitutes a violation of the resolve requirement, in particular of its derives means-end coherence requirement. I believe that these are the cases

philosophers such as Broome had in mind when analysing akrasia. Hence, an agent can be akratic and weak-willed, as I have shown, but in virtue of the violation of two distinct requirements. Also, I believe that it is useful to distinguish those requirements, because they describe failures at different levels of practical agency.

Part II: Individual and social commitments

– What got you here is your word and your reputation.
With that alone, you've still got an open line to New York.
Without it, you're done.
Brother Mouzone to Avon Barksdale - *The Wire*, Season 3, Episode 11

[...] a promise made is a debt unpaid, [...]
The Cremation of Sam McGee, by Robert W. Service

In the previous Sections, I have analysed the elements of practical commitments, and the normative requirements constraining the relation among these elements. The basic picture I have drawn goes as follows. An agent considers certain facts as reasons that count in favour of attaining certain goal or performing certain action, φ. She engages in a practical reasoning process and concludes a normative judgement concerning φ: she decides *that* it is good to do φ; that it is better to do φ than to do ψ; that she ought to do φ. She decides *to* do φ; by choosing, she forms an intention. When the time comes, she does φ. This is how things should work. Many times they do—other times, however, they do not. What could possibly go wrong?

The main problem for rationality is that motivation and judgement do not always go hand in hand. Acting contrary to one's normative judgements, or holding conflicting intentions, is a consequence of this mismatch. Demanding from a rational agent to carry out every intended goal she has is too strict as a requirement, for an agent can change her mind and be rational, and even carry out an intended goal even when there is no reason to

do so, which seems less rational than reconsidering her goal. Hence, rationality has to allow for belief change and reconsideration of one's reasons and goals, as well as for intending and acting in accordance with one's normative judgements, which depend on the reasons the agent considers. On the other hand, as I argued above, the absence of an intentional state does not necessarily entail akrasia. It only does so when the agent intentionally does something else (about which she may not hold any normative judgement) that entails to contradict her previous normative judgement. For example, I believe I ought to finish writing this Section by tomorrow. However, a friend calls me and invites me to go to her birthday party, and I go. I am fully aware that intending to go to a party tonight is going to prevent myself to intend to finish my work by tomorrow. I am intentionally acting against my judgement, and therefore being akratic.

The following schema represents the formulation of the enkratic requirement I suggested in Section 2.2:

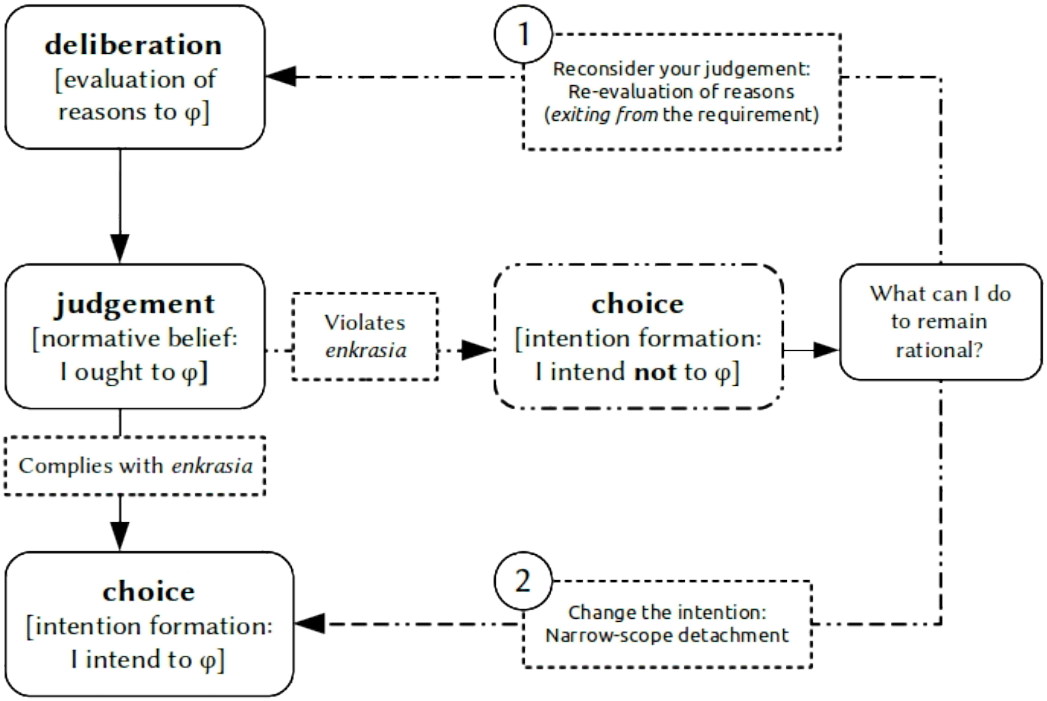

Figure 1: The enkratic requirement

Choosing according to one's normative judgements complies with the enkratic normative requirement. This requirement constrains the agent's practical commitments, because committing herself to do something that she believes she ought not to do makes the agent akratic. Nevertheless, if the agent chooses contrary to her judgements, she can take two possible paths to restore rationality. On the one hand, the agent can exit from the requirement (1) by reconsidering her reasons and changing her judgement. Insofar she does no longer believes that she ought to φ, she is no longer required not no intend not to φ—this is why the agent has exited from the requirement, rather than complying with it. The second option (2) is to change her choice, in order to intend to φ[1]. To put it otherwise: it is presupposed, in rational agency, that the agent has control over intention formation (this is why option 2 is possible), and has rational authority over her deliberation and judgement. This is, it is possible to change our judgements as many times as we want to, as long as the changes in our judgements respond to the reasons we have. Otherwise, it would be a case of self-deception. For example, suppose that Bert judges that he ought to lose weight, because of, let's say, reasons α, β and γ. Bert can reconsider his judgement as many times as he wants, but unless he re-evaluates his reasons, the judgement will be the same. He can find out that reasons α, β and γ do not really support losing weight, or that there are other facts that he was not taking into account, and that are reasons not to lose weight, which may be better or stronger reasons than the reasons he was taking into account before. But if none of this happens, and Bert evaluates the facts just as before, he cannot change his judgement. In this sense, there are similarities between reasons for action and evidence (i.e. reasons for belief). I believe that something is the case in the light of the evidence I have; if the set of available evidence remains invariable, and I have not re-evaluated the evidence I already had, then there is no reason for me to change my belief, even if I feel really motivated to do so.

An agent cannot create reasons only by changing her motivational attitudes towards a fact, for that would be a case of bootstrapping (§2.1.3). But an agent can always

1 There is a third possibility, which would be suspending choice; as I argued above, the absence of intention does not violate enkrasia. I will leave this possibility aside for the sake of simplicity.

revise her reasons, re-evaluate them in the light of new values and facts, and conclude a different normative judgement. Agents, in this sense, are normatively autonomous: they can acquire practical commitments, and also abandon them. Once they judge that they ought to do something, this belief restricts her future choices, but there is room for exiting from the requirement, so it no longer applies to her. The resolve requirement displays a similar normative path. The following figure shows how the resolve and the enkratic requirements interact:

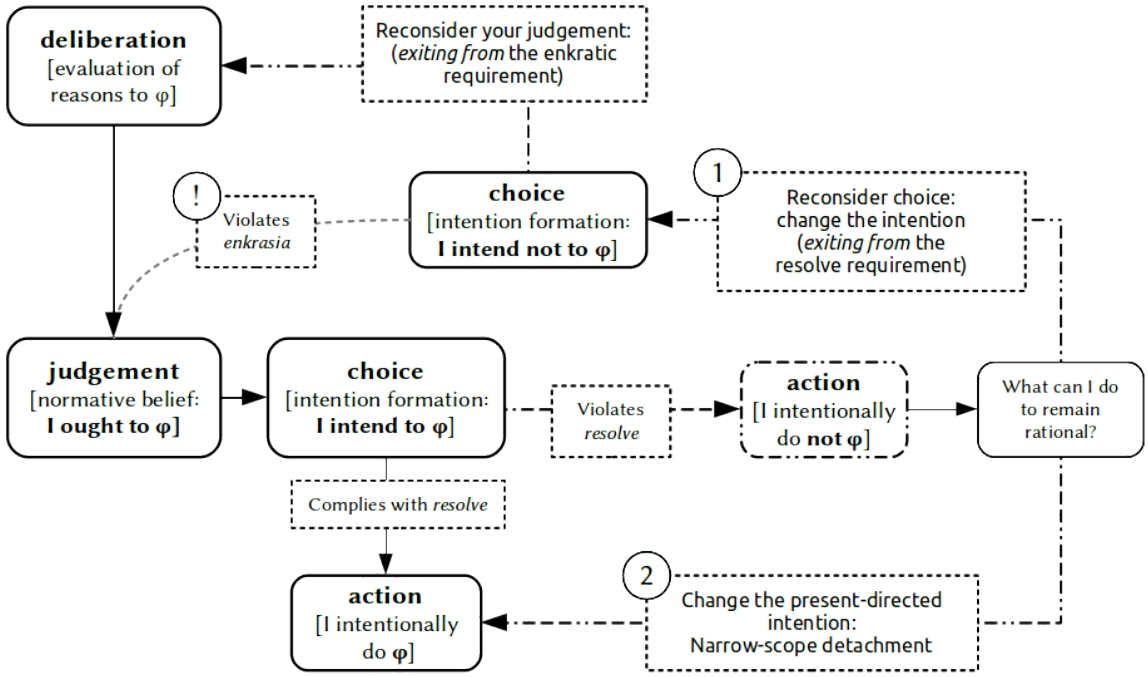

Figure 2: The resolve and the enkratic requirements

Intending to φ requires to exert control over our actions in order to actually do φ. If an agent find herself in a situation or irrationality, she can either change her intentions in order to intentionally doing φ (2), or reconsider her choice and drop her intention (1). However, changing one's mind is subject to the enkratic requirement (!): if you believe you ought to φ, and change your mind and intend not to φ, you are required to revise your judgement in order to avoid akrasia.

My focus in the next Sections will be on the authority to change one's mind. Rationality allows to you revise your reasons for action: you have authority over your deliberation. It is true that, as Holton defends, that reconsidering one's intentions in the absence of a good reason to do so can be regarded as weak-willed; but, as I have argued in Section 1.2.3, there is nothing wrong in reconsidering your intentions as long as this reconsideration does not jeopardise a previous goal of the agent, or contravenes a normative judgement. The agent's evaluative mechanisms change over time, and they also change when certain options are immediately available: temptations are a good example of this. An agent always have the rational authority to revise her judgements. She can do so for what she considers a good reason (there is new evidence available, which is likely to affect her judgement) or a bad reason (she faces a temptation). But, in any case, she is entitled to evaluate the facts available to her and to judge whether they count in favour, or

against, performing certain action, or attaining certain goal. This is why self-directed commands are normatively puzzling (§2.1.3): one and the same agent has the authority to command something to herself, but also to revise her reasons for doing what is commanded, and relieve herself from this command in case her judgement changes. The capacity to relieve oneself from whatever self-imposed obligations is a reason to consider that the Autonomy Thesis has a normative flaw. However, when commitments involve more than one agent (for instance, in the case of a request) the normative structure remains the same as in the case of individual commitments, but only one of the agents, the requester, has the authority to revise the request, and to revoke it in case she believes there are good reasons to do so. The requested agent can either fulfil, violate or ask for a cancellation of, the request, but does not have the authority to revoke it. She has transferred her authority to reconsideration to the requester. Furthermore, both the agent that makes the request and the one who accepts it acknowledge that the fact of accepting that request is a normative reason for the agent to do what is requested. The both accept, usually tacitly, that once the request has been accepted, the agent ought to do what has been requested. If this was a case of self-command, or self-request, the agent would have the authority to revise that judgement, for example, if new information comes out, and change her mind about what she ought to do. In the social case, only the requester has the authority to do this. But if the requester does not release us from the commitment, we are violating it by not intentionally doing what is requested. Thus, the wrong of violating a social commitment, such as a command, a request, a promise or an agreement, is analogous to acting intentionally against a normative judgement: it would be analogous to akrasia. This is the claim I will defend in the following Sections.

The difference between individual and social commitments is not always clear in the literature. For instance, Cohen and Levesque use this story as an example to illustrate what rationality requires from agents in order to be *rational*:

> Some time in the not-so-distant future, you are having trouble with your new household robot. You say "Willie, bring me a beer." The robot replies "OK, boss." Twenty minutes later,

> you screech "Willie, why didn't you bring that beer?" It answers "Well, I intended to get you the beer, but I decided to do something else." Miffed, you send the wise guy back to the manufacturer, complaining about a lack of commitment. After retrofitting, Willie is returned, marked "Model C: The Committed Assistant." Again, you ask Willie to bring a beer. Again, it accedes, replying "Sure thing." Then you ask: "What kind did you buy?" It answers: " Genessee." You say "Never mind." One minute later, Willie trundles over with a Genessee in its gripper. This time, you angrily return Willie for overcommitment.[2]

In this story, it does not seem that Willie is an autonomous agent: it does not have the right to revise its intentions and judgements. Willie's first attempt to satisfy its owner's request fails: the robot changes its mind. What is (rationally) wrong with this? From Cohen and Levesque's perspective, Willie's intention is not as persistent as it should. But they do not specify whether Willie has a good reason to drop its goal. This distinction is highly relevant for internal commitments, this is, intentions, but it needs further analysis in the case of social commitments. There is nothing wrong with changing one's mind, as long as it is done because of, and not in spite of, having better reasons for not doing so[3]; and there os of course nothing wrong in revising one's reasons for action: if nothing has changed, the judgement will be the same, but we always have the right to double-check. So I believe that, from the point of view of individual commitments, Willie is completely free to change its mind. What upsets its owner is the fact that *Willie does not actually have the right to change its min*d, not even to reconsider whether he ought or not bring its owner a beer. This right has been conferred to its owner, because Willie has accepted his request[4].

The overcommitted Willie Model C also fails to please its owner. He had revoked Willie's order, but Willie went on with its plan. Again, I believe that, if this is not a case of social commitment, there is no reason to brand Willie as irrational. Willie may have understood its owner's "Never mind" as an advice, a reason the owner is giving for dropping its goal; however, rejecting advice is not irrational as long as the advice does not provide, or point out, a good reason to change one's mind—a reason that is acknowledged

2 Cohen and Levesque (1990: 214).

3 See Miller (2003) for a similar critique of Cohen and Levesque's account of persistent goal.

4 It seems more a command than a request, but commands involve a threatening aspect that I will not discuss now; so I will keep it as a request, which is simpler to analyse.

by the advised agent. However, the owner has the right to release Willie from its obligations, and sees in Willie's action a failure to understand this release[5]: this is why he brands Willie as irrational.

My aim in the following Sections is to clarify the relation between rationality requirements, normative authority, and the obligations and rights that emerge from a social commitment. The focus of my explanation will be the notion of *expectation*. Two kinds of belief about the others' behaviour[6] are embedded in the concept of expectation. On the one hand, expectations are a form of prediction: they are judgements about what someone will do. These expectations are empirical[7], and rely on background assumptions and beliefs, such as what people use to do under given circumstances, or whether there is a convention or a social norm governing that situation. Also, empirical expectations can be based on other circumstantial cues, such as inferring what someone will do based on what we would do ourselves. On the other hand, expectations can be normative: they also refer to beliefs about what others should do[8]. These expectations also depend on background beliefs, such as the moral or legal norms applied to the situation, or the normative reasons that the agent is expected to have and act upon.

These two dimensions of expectations have been one of the main controversial points amongst theories concerning social norms. In fact, the concept of "norm" itself is

5 An alternative interpretation could be that, instead of releasing it, Willie's owner has in fact changed his request, or command, to the opposite request or command: not only Willie is not obliged to bring him the beer, but is also required not to do it.

6 Of course, it is possible to expect that oneself will or should do A. An agent believes she will do A when she forms the intention to do A, or when she infers that it is likely that she will do A in the future (although she does not intend to do A in that moment). I think this last possibility is closer to a social conception of expectation: it consists in treating oneself as a different agent, and infer what will be her motivations and normative reasons to do something in the future. In this sense, prediction of one's own behaviour can be analysed as if it was the prediction of someone else's behaviour. Hereafter, I will refer to the concept of expectations as entailing a social dimension.

7 Bicchieri (2006); Bicchieri and Xiao (2008).

8 Bicchieri (2006) argues that normative expectations consist in what we think other believe we should do. She refers to beliefs about what others should do as "normative beliefs". This reflexive approach is interesting, as it focuses on how knowing that a behaviour is normatively expected from us is a modifier of our conduct. However,in what follows, I will use Sudgen's terminology, because I will not focus on the motivational power of the beliefs about what is normatively expected from us, but on the relation between normative judgements and normative expectations. See Sugden (1998); (2000b).

ambiguous: it denotes, on the one hand, the regular, typical (i.e. normal) behaviour; on the other, it also refers to the behaviour that is required by some normative standards, such as moral or legal ones. The normative dimension of social norms deals with obligation, avoidance of punishment, feelings of guilt and shame, attributions of moral blame and praise, beliefs about the correctness or incorrectness of someone's actions, and a sense of approval or disapproval of someone's behaviour [9]. The role of normative expectations and beliefs in the agent's behaviour is controversial. Similarly to normative or justificatory reasons, whose relation to motivational reasons is not straightforward, it is not clear why the belief that one should conform to a norm would affect one's decision to conform to that norm. This is the reason why emotional foundations of moral obligation are widespread in the economic literature. From a sociological perspective, the internalisation of norms constitutes the basis for a preference to comply with the norm without the explicit existence of external constraints such as guilt or punishment[10].

The structure of this second part is as follows. Chapter 3 is devoted to the *strategic problem of social commitments*. This problem consists in the need for an explanation of the empirical expectations that arise in a social commitment. In fact, social commitments are usually not only successfully created (i.e. the requester believes that the requested agent will do as requested), but frequently fulfilled—and these facts are puzzling from the perspective of rational choice theory, as well as for the evolutionary approaches. The problem is the following. Commitments are needed when there is an incentive to free-ride (§3.1). Thus, it has to be explained why commitments are credible in the first place, given that it is not in the interest of the agent to fulfil them; and second, once they have been successfully created, why people tend to fulfil them, insofar the incentives to free-ride and the incentives of breaking the promise are the same. Social commitments, thus, entail a change in the payoff structure, and so it becomes in the self-interest of the agent to fulfil them. In order to account for the capacity to modify payoffs, two control mechanisms

9 McAdams and Rasmusen (2006).

10 Vanberg (2008) labels the first kind of approaches "expectation-based", opposed to "commitment-based" approaches, which would focus on the intrinsic motivational force of normative beliefs.

have been proposed: reputation and emotions (§3.2). The capacity of self-control, as I explained in §1.2.1, has been evolutionarily shaped, and these two mechanisms explain why people exert this control over their actions in order to modify their payoff structure, for instance, by making a promise.

The aim of Chapter 4 is to address the *normative problem of social commitments*. The problem here is to explain how an agent can freely undertake an obligation towards another agent. If the strategic problem of commitments asks why people keep their promises, the normative problem wonders why they *ought* to keep them. I first analyse three possible answers to this question (§4.1): practice-based, expectation-based, and reasons-based. In Section 4.2, I argue that the reasons-based account offers a better understanding of the normativity of social commitments. Besides moral or legal obligations, I will argue, social commitments give rise to a set of rights an obligations, that are based on the rational authority an agent has over her reasons for action. Committing oneself entails the exercise of one's normative powers in order to confer another agent certain kind of authority over the justification of one's actions. I will argue that this authority is given through the agreement on the claim that the debtor ought to do what she is committed to, precisely *because* of her commitment. I will ague that social commitments are normative, exclusionary and objective reasons for doing what the agent is committed to.

Chapter 5 is analyses the relation between social commitments and attributions of responsibility. I will argue that the expectations entailed in social commitments serve as a basis for responsibility attributions. I will first present the scope of responsibility (§5.1), in which two main distinctions can be made: retrospective and prospective responsibility, on the one hand, and attributability and accountability, on the other. Then, two criteria for attributing retrospective responsibility (in the sense of *attributability*) will be explored (§5.2): agential capabilities and causal effectiveness. In the third Section (§5.3), I will then argue that both empirical and normative expectations play an important role in determining the causal role of the agent in the production of the outcome. Lastly, the

relation between justification and explanation will be addressed, through the analysis of how excuses and exemptions affect our evaluative judgements of responsibility.

Chapter 3. Empirical expectations

In his famous article "An Essay on Bargaining"[1], Thomas Schelling introduced the concept of commitment into the game-theoretical framework. Committed behaviour, following Schelling, is a kind of strategic action whose goal is to modify the other players' strategies, through the manipulation of their expectations[2]. The two paradigmatic cases of social commitment would be promises and threats. Thus, for a commitment (either a promise or a threat) to be effective, it has to be both believable and believed by the other players. Schelling argues that credibility can be attained by the agent by means of different mechanisms, such as voluntarily ruling out one or more of the available options (either by making some choices impossible or by raising its cost), the power of reputation, and the ability to bargain. Nevertheless, as Schelling points out, credible promises and threats seem to lead to a paradoxical situation: once the commitment has been stated, and having already manipulated the other agent's choices and actions, what's the reason to keep it? Let's imagine, for example, that two agents, Sally and Bob, are playing a public goods game; Sally promises to invest all her tokens into the public good if Bob does the same. Bob believes that Sally's promise is credible, and invests all his tokens. Then, why would Sally keep her promise? From the point of view of narrow strategic rationality, Sally's best option is to free-ride, not only before she made the promise, but also after (and probably she would be more tempted to do so after Bob invested all his tokens in the public good). This problem can be stated as follows: the incentives to cooperate are exactly the same than the incentives to keep a promise or a threat. Nonetheless, experimental results show a

1 Schelling (1956).
2 Schelling (1960: 122).

significant tendency to act according to one's commitments, as well as a tendency to consider the other's commitments credible[3]. It would be unrealistic to state that everyone lives up to their commitments, but the puzzling results do not have to do with the amount of promises and threats that are kept or broken; they are puzzling because, everything else being equal, keeping a commitment is irrational if the behaviour promised is irrational as well—and if it is not irrational, because it is the agent's preferred choice, then the commitment is unnecessary. This Chapter deals with the *strategic* problem of commitment, as stated by Schelling.

3 Ostrom, Walker, and Gardner (1992); Kerr and Kaufman-Gilliland (1994); Kerr et al. (1997); Kurzban et al. (2001); Boadway, Song, and Tremblay (2007); Balliet (2010).

3.1. WHY DO PEOPLE KEEP THEIR PROMISES?

A strategic commitment is a social interaction whose goal is to modify the agent's behaviour through the manipulation of her empirical expectations. From a strategical point of view, then, commitment can only arise in situations in which there is an incentive to free-ride or to cheat. Otherwise, a promise or a threat would not be such but a declaration of intentions: "a promise or a threat must be to do something that the individual would not be otherwise motivated to do. That is what distinguishes these pledges from mere forecasts"[4]. Thus, commitment serves as a control mechanisms in order to incentive cooperation and overcoming the temptation of free-riding: "commitment is a means by which player can assure one another that they are not going to free ride on other's contributions, so that group members can contribute without fearing that they will be free ridden"[5]. However, it is not clear what mechanisms promote commitment in the first place: when there is a temptation to free-ride, there is indeed the same temptation to propose a false commitment. In this section, I will explore the problems of commitment in strategic contexts. Commitment works as a mechanism that enhances cooperation in social dilemmas, but its credibility and effectiveness depends on other mechanisms that enable pro-social behaviour.

Thus, how does a commitment modify the payoff structure? What other incentives, besides the explicit payoffs, are taken into account by the agents? Is the temptation to free-ride in a game the same kind of temptation as the one that leads to deception? Or, stated otherwise, if an agent is afraid of cooperating because she believes that the other agents are going to free-ride, then she would also have reasons for believing that, in the case that all the other players commit themselves to cooperate, they are also going to cheat, and break their commitments. Following this argument, Sánchez-Cuenca (1998) claims that there is a trade-off between the need of a commitment and its credibility:

4 Hirshleifer (2001: 309).

5 Kurzban et al. (2001: 1663).

> A commitment is credible when everyone expects that the person who makes it cannot renege on it. But it happens that the conditions that make it difficult to renege are the same conditions that make it difficult to commit. Thus, the more credible a commitment is, the more unlikely that the commitment can be made. [6]

Hence, the more a commitment is needed, the less likely this commitment will help in solving the problem. Commitments are needed in cases in which the payoff structure incentives the player to free-ride; and making a commitment is not going to change this situation. When the temptation to free-ride is stronger, the more needed a commitment is, but the less credible it will be, so the less likely it will solve your problem. The empirical problem of commitment is thus the following[7]: how does a commitment modify the agent's incentives, constraining their choices? And why are commitments fulfilled in the absence of an external enforcing mechanism? Some authors argue that a social commitment necessarily entails an external enforcing mechanisms, which can be self-imposed, such as restricting one's options or making them more costly[8]. These manipulations of an agent's set of options in order to make a commitment credible are called commitment technologies. Once the commitment technology is set up, there is no need for an additional mechanism to incentive its fulfilment, because fulfilling it matches the agent's self-interest. However, I will argue that it is possible to establish credible commitments without necessarily restricting one's options.

Human interactions can greatly vary in complexity, so there is a wide scope of situations in which agents interact. The variety of actions involved in a social interaction makes complicated the task of choosing the criteria to build a typology with; this is why, depending on the goal of the typology, different criteria tend to overlap[9]. However, there are two important classifications of social interactions.

The first of them is Hamilton's classification[10]. This classification of social interactions considers a basic situation in which an agent performs an action and another

6 Sánchez-Cuenca (1998: 86).
7 Frank (2003).
8 See Elster (2000); Elster (2003); see also Chapter 2.
9 Álvarez (2006).
10 Hamilton (1964).

agent is affected either by the consequences of that action, or by the action itself. Hamilton claims that there are four basic kinds of social interaction, based on how the well-being of each individual is affected by the action:

		Effect on recipient	
		+	-
Effect on actor	+	Mutual Benefit	Selfishness
	-	Altruism	Spite

Table 3: Hamilton's matrix, adapted by West, Griffin and Gardner [11]

Second, from the point of view of game theory, a social interaction is a situation in which individuals are affected by the choices of other agents[12]. The classification of different interactions depends on the context and the agent's interests, this is, in how the payoffs for each player are structured. In the game-theoretical framework, games are classified following the degrees of conflict and coordination expected from the players, depending on the structure of the payoffs. It is necessary to point out that the concept of coordination used in game theory differs from the concept of coordination that I have analysed above. A game of coordination is a game in which it is possible to reach an agreement about the individual choices, because players are interested in knowing what the other players will do, and also in letting the others know what the agent is going to choose: "Coordination problems are often viewed as simple to solve. In large part this is because actors have similar interests, and, although they may not care about which solution is imposed, they all agree that some solution is needed"[13]. There are four basic kinds of games, which correspond to four kinds of social interactions: cases of pure common-interest (in which pure coordination is expected), Battle of the Sexes games, Prisoner's Dilemma games, and inessential games (which are pure conflict situations)[14].

11 West, Griffin, and Gardner (2007: 418).
12 Bramoullé (2007).
13 Wilson and Rhodes (1997: 767).
14 Parisi (2000).

3.1.1. Pro-sociality and altruistic behaviour

The concept of pro-social behaviour covers a broad category of interactions, which include cooperation, helping others, sharing resources, and altruistic actions [15]. Its analysis takes into account cognitive, biological, motivational and social processes [16]. Despite being a quite common phenomena, pro-social behaviour challenges some central assumptions of the evolutionary and economic theories of strategic interaction, because these models claim that the goal of social interactions should be fitness, or utility, maximization. Thus, cooperation is expected only in those cases in which both agents are better off through cooperation. In other situations, an agent would increase her fitness or her welfare by free-ridding on the others: without assuming any cost, she can benefit from the others' actions.

The capacity for making credible commitments has been understood as a mechanism that enables and promotes pro-social behaviour: knowing that others will cooperate enhances the rate of cooperation within the group. Promises, threats, agreements and contracts are commitment technologies that enforce cooperative behaviour. However, they inherit the problems concerning pro-sociality: once the commitment is effective there is no reason to fulfil it. When the other agents have cooperated believing in the honesty of the commitment, the temptation to free-ride does not disappear, unless other mechanisms intervene. Furthermore, knowing that free-ridding is the best strategy of the agent, her commitment should not be credible at all. However, despite this theoretical problems, people make and fulfil credible promises. This conflict between our theories of rationality and the observed behaviour is called the puzzle of pro-sociality:

> Individuals often do better by coordinating and sharing the benefits of their activities rather than each acting alone. The benefit accruing to the group from each individual's cooperation in such cases is greater than the cost to the individual, but nonetheless, each individual would be better off not incurring the costs of cooperation, and simply benefiting from the efforts of the other group members. If all participants follow this self-interested logic, however, cooperation will fail. When it is maintained, cooperation is altruistic, in the sense of being

15 Dovidio and Penner (2004).

16 Penner et al. (2005).

> group-beneficial but personally costly. Why are such altruistic behaviours not driven out by self-interested agents? This is the puzzle of pro-sociality. [17]

Since the 1960s, the claim that natural selection operates exclusively at an individual level has been challenged. The tension between the individual and the collective level, and the role of pro-social behaviour in the evolutionary history of species, however, are still controversial. The debates mainly discuss whether it is possible to properly speak of a group-level natural selection that favours individual pro-social behaviour, or there are individual and non-immediate adaptive advantages for those individuals who prefer to cooperate rather than to free-ride[18].

Cooperative interactions are shown in the left column of Hamilton's matrix of social interactions (see Table 3). To cooperate means to act in a way that benefits the recipient of the action, and choosing to do so precisely because of its beneficial consequences on the recipient[19]. Cooperation can be either beneficial for the actor, thus generating a situation of mutual benefit, or it can be costly, and therefore considered a case of altruism. Then, the paradox of pro-social behaviour is the following: in situations in which cooperation is costly, there is an incentive to free-ride; however, if every individual was a free-rider, then cooperation would no be possible and the final result would not be beneficial for any individual.

The explanations of cooperation use two strategies to approach the paradox. The first of them consists in denying that Hamilton's matrix is an adequate model of social behaviour, because it only considers the individual level, and ignores the group benefits derived from cooperation. Wilson[20] argues that natural selection does not only operate at an individual level, but also at a group level: groups with better proportion of altruistic individuals perform better, and thus obtain better adaptive results:

> Altruism is selected against at the individual level because non-altruists have the highest fitness within all mixed groups. Altruism is favored at the group level, however, because

17 Gintis (2003: 157).

18 Penner et al. (2005).

19 West, Griffin, and Gardner (2007: 416).

20 Wilson (1975); for an overview of group selection theories, see Wilson and Sober (1994).

group fitness is directly proportional to the frequency of altruists in the group [21].

The theory of group selection has been criticised for different reasons. For instance, Nesse[22] argues that the works on group selection do not deal with a central conceptual problem: the existence of traits that are adaptively beneficial at the group level, but that are nonetheless prejudicial at the individual level. On the other hand, it is argued that other competing theories are more broadly applicable[23]. Despite its critics, the theory of group selection has not been abandoned, but its application is limited to human groups, in which survival does not merely depend on natural selection, but also in cultural selection. There is empirical support to the claim that social norms and institutions could be the result of cultural selection mechanisms[24]. However, as Fehr and Fishbacher[25] point out, it is needed to introduced additional mechanisms, such as altruistic punishment, for these norms and institutions to arise in the first place.

The second kind of approaches to the paradox of altruistic behaviour focus on the individual benefits of cooperation and altruism, which would not really imply costly cooperation, but a cooperation whose benefits are not immediate. There are two main theories sharing this perspective. The first of them is the theory of inclusive fitness or kin selection, proposed by Hamilton [26]. This theory focuses on the tendency of individuals that share genetic information to mutually benefit each other, thus facilitating their reproductive success, and raising the probability to pass their genes to the next generation.

On the other hand, the theory of reciprocal altruism, introduced by Trivers [27] aims to explain cooperative behaviour between individuals that are not genetically related, and applies a game-theoretical framework to scenarios in which there is an incentive to free-ride, such as the Prisoner's Dilemma (see Table 4.2). From this point of view, many

21 Wilson and Sober (1994: 591).
22 Nesse (1994).
23 West, Griffin, and Gardner (2008).
24 van den Bergh and Gowdy (2009).
25 Fehr and Fischbacher (2003: 789).
26 Hamilton (1964).
27 Trivers (1971).

behaviours, previously considered altruistic, are not indeed disinterested. When punishment mechanisms are introduced, cooperating is an investment for the agent, warranting future interactions: "The idea here is that individuals can take turns in helping each other, for example by preferentially aiding others who have helped them in the past. Trivers termed this 'reciprocal altruism'"[28].

	Cooperate	Defect
Cooperate	3,3	0,5
Defect	5,0	1,1

Table 4: The Prisoner's Dilemma payoff matrix [29]

Theses analysis pay attention both to the general tendency of cooperation within the group, and to the interaction between two individuals in the group: the latter is supposed to explain the former. However, large groups are problematic, because the larger the group, the less advantageous is to set up a control mechanism to avoid the temptation of free-riding. In small groups, individuals tend to interact repeatedly with other members. Thus, the outcome of previous interactions can be recorded and used to decide whether to engage in a new interaction with the same individual –there are not one-shot encounters, but repeated interactions. On the contrary, within large groups, the probability of repeating an encounter diminishes, and thus additional mechanisms to keep the rate of cooperation are needed.

Inspired by Trivers' theory of reciprocal altruism, Axelrod and Hamilton[30] tackled the problem of what strategies are evolutionarily stable -this is, a strategy such as a spontaneous apparition of a mutation of that strategy do not alter its initial predominance- in the Prisoner's Dilemma game. They showed that the "tit-for-tat" strategy turned out to be stable, robust, and plausible to appear for the first time in a randomized system. In an iterated Prisoner's Dilemma, the "tit-for-tat" strategy consists in

28 West, Griffin, and Gardner (2007: 240).
29 Axelrod (1980: 5).
30 Axelrod and Hamilton (1981).

starting the game by cooperating, and then copy the other player's last movement: it is a strategy based on reciprocity, and has become a paradigmatic explanation of reciprocal altruism [31]. There are other similar strategies that are also able to punish defection, such as 'always cooperate and punish your partner after each round in which it failed to cooperate as well', or 'play tit-for-tat and in addition punish the partner for each defection' or 'start cooperatively, punish your partner the first time it fails to cooperate and switch to defection if the punishment does not alter the partner's behaviour'[32]. These strategies have in common that the player is sensitive to the other player's previous choices, and thus choice is not exclusively based on the immediate payoffs of the encounter, but includes external considerations.

The concept of strong reciprocity is central to this theoretical framework. It refers to the predisposition to cooperate with others, and to punish defective agents, even in cases in which this behaviour cannot be justified through self-interest, kin or reciprocal altruism[33]. This strategy is a combination of various control mechanisms that incentive cooperation:

> Strong reciprocity is a combination of altruistic rewarding, which is a predisposition to reward others for cooperative, norm-abiding behaviours, and altruistic punishment, which is a propensity to impose sanctions on others for norm violations. Strong reciprocators bear the cost of rewarding or punishing even if they gain no individual economic benefit whatsoever from their acts. In contrast, reciprocal altruists, as they have been defined in the biological literature, reward and punish only if this is in their long-term self-interest [34].

The difference between reciprocal altruism and strong reciprocity lies in that a reciprocal altruist will only cooperate if there are future returns for cooperation, while a strong reciprocator will respond to the kindness perceived in the other player, rather than in the immediate or future payoffs of the game. Strong reciprocity is observed to take place both in real interactions and in laboratory experiment [35], and plays a central role in the

31 Nowak and Sigmund (1993).
32 See Bshary and Bergmuller (2008); see also Hammerstein (2003); Nowak (2006).
33 Gintis (2000b).
34 Fehr and Fischbacher (2003: 785).
35 Fehr, Fischbacher, and Gächter (2002).

enforcement and content of social norms[36], specially when third-party agents are allowed to reward or punish the agents involved in an interaction.

In brief, control mechanisms favour that, in repeated encounters, agents increase their tendency to cooperate; nonetheless, many of these control mechanisms are costly. Then, what is the incentive to set up a control mechanisms in a one-shot encounter? This is, why to punish a cheater, incurring into costs, if the agent will not interact again with the cheater in the future? Experimental evidence shows that, in larger groups, the level of cooperation decreases[37]. Altruistic punishment and strong reciprocity have been proved to be effective mechanisms to maintain the rate of cooperation within larger groups[38], which enhances the group fitness[39]. It is thus necessary to specify the motivational mechanisms underlying this kind of behaviour because, despite of being adaptive at the group level, they do not offer immediate advantages for cooperative agents.

A different way of dealing with the problem of commitment consists in challenging the relation between commitment and self-interest. The rationality of committed behaviour is problematic because the benefits of this behaviour are not immediate, or even non-existent (for example, in the case of altruistic punishment to an agent with which there will not be a repeated interaction). Thus, some authors argue that preferences are not exclusively guided by self-interest or welfare maximization, but other factors such as moral considerations or the compliance with social norms are also relevant in the formation of preferences.

3.1.2. Sen on commitment as altruistic motivation

Amartya Sen's "Rational Fools"[40] is nowadays one of the most cited and commented works in the field of rational choice theory (RCT). From Sen's point of view,

36 Fehr and Fischbacher (2004b).
37 Fehr and Fischbacher (2003).
38 Boyd et al. (2003); Fehr and Rockenbach (2004).
39 Gintis (2000a).
40 Sen (1977).

commitment cannot be accommodated in RCT explanations because it opens a wedge between welfare and choice. In "Rational Fools", Sen argued that we must distinguish between two separate concepts: sympathy and commitment. The former corresponds to the case in which the concern for others directly affects one's own welfare: "If the knowledge of tortures of others make you sick, it is a case of sympathy; if it does not make you feel personally worse off, but you think it is wrong and you are ready to do something to stop it, it is a case of commitment" [41]. Later, in "Goals, Commitment and Identity" [42], Sen developed the theoretical distinction between self-centred welfare, self-welfare goal, and self-goal choice, and placed this distinction in the core of RCT models. He argues that sympathy only violates the self-centred welfare condition, because the welfare of others influences our own welfare. RCT can easily explain this kind of "altruism", due to the fact that an agent's welfare increases by making other's welfare increase as well. Sen argues that commitment, however, involves making a choice which violates either the RCT requirement of self-welfare goal or self-goal choice. Sen claims that "commitment is concerned with breaking the tight link between individual welfare (with or without sympathy) and the choice of action (for example, being committed to help remove some misery even though one personally does not suffer from it)"[43].

Sen's critique focuses on the self-interested assumptions of classic RCT[44]. In spite of the attempts to broaden the concept of welfare in order to include altruistic preferences[45], Sen argues that broadening the concept of welfare is not a satisfactory solution, because the underlying problem is the connection between welfare and preferences: Sen claims that an agent is able to choose an option that violates her preferences, because the choice is not exclusively made on the basis of the agent's welfare, but on the agent's commitments.

41 Ibid., 319.
42 Sen (1985).
43 Ibid., 7–8.
44 Debreu (1959).
45 Becker (1974).

Thus, Sen's concept of commitment is different from Schelling's, although they can be both applied to some specific situations. For example, fulfilling a promise can be counter-preferential, in Sen's terms, when it is strategically better to violate it. The claim that agents are able to make counter-preferential choices is highly controversial, because it undermines the common understanding of preferences. Other alternative concepts of preference can broaden the motivational scope of the agent, and thus include committed action in the set of preferred actions. For example, Hausman [46] argues that commitment does not entail counter-preferential choice if preferences are seen as all-things-considered rankings. Rather, Hausman argues, commitment should be invoked as one of the preference formation mechanisms, this is, as a kind of motivation, among others. From a different perspective, but also regarding broader concepts of preferences, it has been argued that individuals do not only reason to maximize their individual welfare, but they also reason as participants of a group or team[47]. Individual agents would have team preferences that are not reducible to individual preferences, and they try to fulfil these preferences when acting as a group member. However, due to its relevance for the analysis of collective action, I will focus on this theoretical approach in Chapter 6, which is devoted to the analysis of collective commitments.

On the other hand, Sen argues that commitment violates the self-goal choice assumption. Commitment consists thus in the adoption of some other agent's goals, and the willingness to promote this goal; threats cannot be considered commitments from this perspective[48]. However, adopting someone else's goal is different from acting to promote someone else's goal. Drawing a distinction between goal modifying and goal displacing, Pettit[49] argues that, while the modification of an agent's goals in order to consider other people's goals is quite common, the possibility of acting in order to attain a goal that the agent does not have is highly implausible: the notion of agency entails a relation between an agent's goals and actions. While goal displacing requires a departure from rational

46 Hausman (2005).

47 Sugden (1993); (2000a).

48 Guerini and Castelfranchi (2007); see Chapter 4 for a further analysis of social commitment as goal sharing.

49 Pettit (2005).

choice theory, goal modifying can be accommodated by modelling the deliberation process that allows agents to include other agent's goals as owns.

I believe that Sen's critique has a normative dimension that cannot be accommodated within the strategic rationality framework. In this sense, RCT is not a theory of action or rationality, but a framework to explore the formal restrictions on the structure of preferences[50]. RCT includes some assumptions about the content of the agents' preferences, but leaves room for motivational theories to explain the formation of preferences.

3.1.3. Socially-mediated preferences

Social norms are one of the most invoked concepts in the social sciences; however, there is not a widely accepted account on how they are formed, and what mechanisms promote their enforcement[51]. Indeed, explanations in terms of social norms are not usual in behavioural economics, partly because their conceptualization is vague and therefore it is problematic to include them in formal models[52]. On the one hand, social norms explicit what kind of behaviour is expected from the agents in different contexts, and therefore enable the prediction of sanctions in case of non-compliance. Social norms are enforced by external mechanisms such as rewards and punishments, and often these two responses to the other's actions are guided by social norms, such as norms of fairness or reciprocity. Thus, norms tend to be self-enforcing: the violation of a norm of reciprocity may be responded with a sanction; and this process of sanctioning is norm-mediated, and not a product of deliberation[53]. As it has been showed in the previous section, the enforcement of social norms is related to strong reciprocity[54]. Not only agents have a tendency to comply with the norm, but they are also willing to sacrifice part of their welfare to reward

50 See Güth and Kliemt (2004); Brennan (2007).
51 Fehr and Fischbacher (2004a).
52 Bicchieri and Chavez (2010).
53 Posner and Rasmusen (1999).
54 Fehr and Fischbacher (2004b).

or punish other agents. The compliance to norms of cooperation and fairness increase when strong reciprocators are able to make credible commitments to punish deviant behaviour[55].

Besides having a preference for acting consistently, agents also tend to follow informal rules of fairness and to avoid inequity. In fact, these two concepts are deeply interlinked:

> [W]e model fairness as self-centered inequity aversion. Inequity aversion means that people resist inequitable outcomes, i. e., they are willing to give up some material payoff to move in the direction of more equitable outcomes. Inequity aversion is self-centered if people do not care per se about inequity that exists among other people but are only interested in the fairness of their own material payoff relative to the payoff of others. [56]

Sometimes norms of fairness conflict with self-interest, understood as a maximization of self-welfare. This is why the tendency to comply with norms of fairness has been referred to in the literature as if the agent chose according to her "social preferences". Models of social preferences assume that people are self-interested, but are also concerned with the payoffs of the other players[57]. For example, in a Dictator's game[58], a rational agent would give zero tokens to the other player; however, on average, people share a 30% of their tokens[59]. This 30% is understood as a measure of social preference. Reciprocal altruism, inequity aversion and strong reciprocity are usually modelled as social preferences. Nonetheless, the explanation of the formation of this kind of preferences is usually left unattended in the economic literature, partially because the answer to the motivations underlying preferences may have an evolutionary (both biological and cultural) origin, leaving its analysis for evolutionary biologists and anthropologists. One possible approach to the formation of social preferences, including altruistic behaviour, is that they could be the result of human 'docility', in Simon's terms[60]. Docility is the "tendency to depend on

55 Sethi and Somanathan (2005).

56 Fehr and Schmidt (1999: 819).

57 Charness and Rabin (2002).

58 The Dictator's game is actually not a game in the strict sense. In this scenario, a player (the dictator) is given an amount of tokens and is given the possibility of offer a share of the tokens to the other player, who has no active role in the game.

59 Croson and Konow (2009).

60 Simon (1990); (1993).

suggestions, recommendations, persuasion, and information obtained through social channels as a major basis of choice"[61]. Human rationality is limited and not able to support optimization, and it is therefore approximate and bounded[62]. Thus, by learning how to appropriately respond to a social scenario (such as sharing resources), the agent employs heuristic mechanisms which, rather that maximizing the outcome, optimize the decision process.

Lastly, it is possible to model the compliance to social norms as a kind of preference to follow the norms under specific conditions, such as the belief that other players will do so, and the belief that other players think that the agent should comply with the norm[63]. Both empirical expectations, which are the agent's beliefs about what the other agents will do, and normative expectations, which are the beliefs of the agent about what other players belief she should do, are determinant for conditional cooperation[64]. Also, the salience of certain social norms in the social setting can affect the decision of following the norm[65]. Thus, the conditional preference for rule-following would, from this theoretical point of view, conditioned by the beliefs of the agent about what other will do, except in the cases of moral norms, which demand an unconditional commitment[66]. In the next section, I will apply the mechanisms reviewed so far to the analysis of social commitments. Promises and threats are supposed to be binding mechanisms that enhance cooperation; however, they inherit the same problems than the kind of behaviour they are suppose to promote, regarding their motivation.

61 Simon (1993: 156).
62 Gigerenzer and Selten (2002).
63 Bicchieri (2006).
64 See Bicchieri and Xiao (2008).
65 See Bicchieri and Chavez (2010).
66 See Bicchieri, Nida-Rümelin, and Spohn (2000).

3.2. Mechanisms that enable credibility and trust

Thus, additional control mechanisms are needed to explain both the fulfilment and the credibility of commitments. Following Nesse [67] there are four reasons to believe that a commitment is likely to be fulfilled. First, a commitment can be self-enforcing: after the creation of the commitment, the action involved becomes the best option for the agent, following her self-interest. In this case, the creation of the commitment implies a restriction of options, either by making them unavailable, or by raising its costs. An example of a self-enforcing commitment would be burning one's bridges or ships[68]. Second, a commitment can be reinforced by external incentives controlled by third parties. For example, a contract is a commitment that is enforced by legal punishment. These two mechanisms turn the fulfilment of a commitment into a self-interested action. Besides the sceptical argument that states that commitment faces the same problems than cooperation, and thus is not effective without external constraints, there is large experimental support to the claim that communication enhances cooperation in social dilemmas[69]. Thus, other internal mechanisms play a role in the explanation of the effectiveness of a commitment. The third mechanism Nesse points out is reputation. Lastly, the binding force of a social commitment can be related to emotions. Nesse claims that, when the third (reputation) and fourth (emotions) mechanisms are into play, the commitment is subjective. The only reason why these commitments are effective is the power of persuading other agents that the committed agent will act against her self-interest, understood as the best option for the agent if the commitment did not take place. On the other hand, experimental results show that it is not there is a complex relation between external and internal mechanisms; for example, when evaluating both whether to accept a commitment (by modifying one's behaviour) or whether to carry out

67 Nesse (2001).
68 Elster (2003).
69 See Balliet (2010) for a meta-analysis of related experiments.

a threat, or fulfil a promise, several factors such as expectations, the possibility of punishment, and the communication that enables the commitments come into play[70].

The mechanisms that enable credibility and the mechanisms that enforce the fulfilment of commitments are different, although deeply related. Credibility would not be possible without a relatively high prevalence of fulfilled commitments, because the agent would expect its violation. It is also important to consider the differences between positive (promises) and negative (threats) commitments[71], specially in the case of subjective commitments, this is, not enforced by external agents or by a modification of the payoffs structure. When a promise is effective, the promisee changes its behaviour and expects that the promiser fulfils his commitment. The mechanisms involved in this fulfilment have to take into account both the consequences of not fulfilling the promise, such as a changes in reputation or status, or involved negative emotions, and the mechanisms that override the temptation to free-ride: if the other agent has modified her behaviour and behaved cooperatively, for instance, then the promiser has an incentive not to fulfil her promise. On the other hand, a threat consists precisely in committing oneself to perform an action that is against her self-interest, in order to disincentive an agent to free-ride, or to punish such a behaviour. But carrying out a threat is costly, and its effects on the threatened are relevant for future interactions, not the present one. Thus, the agent who threatens does not have an incentive to free-ride, but rather she needs for incentives for motivating altruistic punishment.

Reputation is able both to incentive the fulfilment of a threat and its credibility. In cultures of honour, the defence of one's own reputation is achieved through violent and disproportionate threats, which are usually carried out in case when the agent is challenged[72]. Not carrying out a threat can cause that, in future encounters, the agent is free-ridden. Furthermore, the defence of honour is related to social emotions such as rage

70 Ostrom, Walker, and Gardner (1992).
71 Castelfranchi and Guerini (2007).
72 See Cohen and Vandello (2001).

or anger, and shame and humiliation, which are relevant to explain the fulfilment of threats in cases in which it is more advantageous not to do so[73].

Social emotions play an important role in the explanation of the effectiveness of commitments, because they promote credibility and motivate their fulfilment. For example, guilt-aversion can incentive the avoidance of breaking a commitment[74]. The social emotions that mediate cooperation also mediate the effectiveness of a commitment. The detection of the other player's emotions are key in order to assign credibility to a commitment[75]. Vanberg[76] argues that emotions create an indirect bond between preferences and promises, because they are able to modify the second-order beliefs of the agent.

Regarding positive commitments, trust is a necessary mechanism for assigning credibility. Trust is a complex motivational and cognitive state that enables the generation of empirical and normative expectations about the other's behaviour under risk circumstances[77]. The effectiveness of a social commitment depends on the successful manipulation of the other agent's choices, and a minimum level of trust is required between the two agents for a commitment to be credible, and hence effective [78]. A relation of trust entails a disposition to rely on the other agent for the fulfilment of one's own goals[79]. Thus, trust leads to the credibility of commitments when there is a situation of dependence and uncertainty[80]. By trusting the other, the agent incurs in costs: she chooses according to a future payoff, rather than a present one. Without uncertainty, trust is not necessary, because the agent expects the choices of the other player independently of the commitments in which that player has incurred.

73 See Mosquera, Manstead, and Fischer (2002).
74 See Battigalli and Dufwenberg (2007); Ellingsen et al. (2010); Charness and Dufwenberg (2010).
75 See Frank (2001); Irons (2001).
76 Vanberg (2008).
77 Origgi (2008).
78 See Hardin (2003); Simpson (2007).
79 Castelfranchi and Falcone (1999); (2002); see also Chapter 5
80 Barbalet (2009)

Lastly, the fulfilment of a commitment can be motivated by social preferences of the agent, or her preference for consistent behaviour. These social preferences are conditioned to the expectations of the agent of other agents' compliance to social norms. Promises and contracts are usually regulated by social norms that dictate the conditions under which it is expectable from agents to fulfil their part of the agreement, and thus condition their credibility.

Communication and reputation

The problems of pro-social behaviour stated in the previous section can be extrapolated to the analysis of communication. Truthful communication has the same problems as cooperative behaviour: additional mechanisms are needed to overcome the temptation of sending wrong signals to take advantage.

Communication is a necessary part of a social interaction, in which the actions of an individual generate a signal that modifies the behaviour of the receiver[81]. There are two main theoretical approaches to animal communication. The first of them considers communication as a mechanism to transmit information, and has its theoretical origins in Darwin's work. From the point of view of group selection, clear and non ambiguous signals are evolutionarily advantageous, specially when they are meant to inform about states of affairs that have not been directly experienced. To overcome the temptation of emitting false signals, and free-ride on the honest signals of others, Zahavi[82] proposes the existence of a "handicap principle". According to this principle, the communication of honest signals arises when the cost of sending the signal is elevated, and therefore cheating becomes too costly.

The second approach to communication comes from sociobiology, and it criticises the claim that the function of communication is to share information: communication would be better understood as the manipulation of the other's behaviour: "the evolution

81 Wiley (1983)

82 (1975); Zahavi et al. (1999)

of many animal signals is best seen as an interplay between mind-reading, and manipulation"[83]. A cheater would have adaptive advantage within a group in which individuals always send honest signals, thus complete honesty cannot be an evolutionarily stable strategy. Krebs and Dawkins suggest that the goal of sending a signal is to manipulate the receiver in a way that it fulfils the sender's self-interest. The receiver needs then to predict what action the sender will perform.

Maynard Smith and Harper[84] suggest a combination between the informational and the manipulative approaches. They argue that it is not evolutionarily stable for the receiver to modify his behaviour (just as the sender pretends) unless the information contained in the message is credible and useful. For example, a threat signal will not have any effect on the receiver if he does not identify that signal as a credible threat. Game-theoretical models of signal credibility confirm the inverse correlation between the cost of a signal and the incentives to free-ride by sending that signal[85]. Besides the cost of producing a signal (intrinsic cost), reputation also raises the cost of sending false signals.

Reputation affects the willingness of individuals to engage in a repeated social interaction with another individual. Triver's reciprocal altruism consists in cooperating with those who have been cooperative in earlier interactions; Alexander's[86] strategy of indirect reciprocity consists in cooperating with those who either have been cooperative in earlier rounds, or that it is known by the agents that they have been cooperative in earlier rounds. Strategies based on reputation, such as reciprocal altruism or indirect reciprocity, need a mechanism to register the past behaviour of individuals, and a set of rules to assess how to behave, depending on the information available about the other individual[87]. It is possible to distinguish different levels of complexity in the mechanisms that generate a reputation system. The first of them is based on the emotions of fear and submission[88].

83 Krebs and Dawkins (1984: 380).
84 Smith and Harper (1995); (2003).
85 Gintis, Smith, and Bowles (2001).
86 Alexander (1987).
87 Nowak and Sigmund (2005).
88 Henrich and Gil-White (2001).

The mechanism that promotes (or restrains) cooperation is the set of emotions that the other individual causes on the agent, and these emotions can be prompted by previous interactions, or by the observation of interactions. A more complex level would involve more complex cognitive processes, such as the possibility of making predictions about the behaviour of others. In this level, instead of manipulating directly the behaviour of others, the agent tries to manipulate their expectations, thus generating trust relations. This kind of reputation is exclusively found in human societies. The problem of the credibility of signals and its relation to reputation switches the focus: instead of asking why agents send honest signals in contexts in which not doing so enhances the individual's fitness, the question would be why agents send honest signals in contexts in which not doing so is strategically advantageous. I will return to this point in the second part of this chapter; now, I will explore the role of emotions in the promotion of pro-social behaviour.

The role of emotions

Evolutionary analysis show that emotions have survival and reproductive functions, which are manifested in four different levels: intra-individual, dyadic, group, and cultural[89]. While the functions of emotions in the first level tend to enhance the individual fitness, the same functions in the other three levels usually favours the creation of social bounds and cooperation.

From the point of view of strategic rationality, one-shot encounters are essentially different from repeated encounters. In a one-shot game, control mechanisms such as long-term investments or the building of a reputation cannot arise, because the agent would lack of the motivation to do so. Social emotions play this role: they motivate cooperation, serve as a guide to choose a partner for interaction, and enable the creation and perdurance of long-term relationships[90]. Furthermore, detecting the other's emotions also serves as a mechanism for evaluating the interaction partners and avoid cheaters or

89 Keltner, Haidt, and Shiota (2006).

90 See Frank (2001); Gonzaga et al. (2001); Back and Flache (2008).

free-riders[91]. The feeling of anger or frustration after being cheated disincentive future interactions with the same individual, and can motivate altruistic punishment. Fehr and Gächter[92] argue that in a Public Goods game[93], cooperation only arises when agents have the possibility of punishing free-riders. Their study shows that there is a correlation between the intensity of the emotion felt and the punishment executed. In the Public Goods game, those players who have invested more tokens report the most intense negative emotions, and this intensity also increases when the amount of tokens invested by the other player is lower. Other studies point out the necessity of including the role of social norms to understand the relation between emotions and expectations of the agents: expectations are based on the fulfilment or violation of the agent's expectations, and these, in turn, are generated following social standards[94]. In what respects to the differences between one-shot and repeated encounters, experimental evidence shows that the knowledge of the kind of encounter the subject is involved affects the strength of the emotions that motivate punishment varies, but does not disappear in one-shot encounters[95]. Lastly, emotions do not only play a role in the motivation of punishment, but the expression of a negative emotion serves as a punishment mechanism through the generation of feelings of guilt or shame[96].

In brief, communication plays a central role in the generation of expectations on other agents. Individuals thus hold a reputation based not only on their willingness to cooperate, but also on the honesty of the signals they send. On the other hand, social emotions serve as incentives to cooperate or defect to other agents depending on their previous and present conduct.

91 Cosmides and Tooby (2004).

92 Fehr and Gächter (2002).

93 A Public Goods game is an example of social dilemma. The players have the chance to invest an amount of tokens in the production of a public good; the tokens invested in the public good are multiplied and distributed equally among the players, even among those who have invested zero tokens (free-riders).

94 See Hoffman et al. (1994); Bosman and Van Winden (2002); Wu et al. (2009).

95 Fehr and Henrich (2003).

96 Xiao and Houser (2005).

To sum up, the success of strategic commitments lies in other mechanisms that enable pro-social and cooperative behaviour. Particularly, reputation and emotions make commitments credible, and contribute to their fulfilment. On the one hand, because of reputation systems, making credible commitments and fulfilling them is in the agent's self-interest. On the other hand, social emotions promote punishment and rewarding behaviour, which also affect the agent's behaviour. Thus, the solution to the puzzle of strategic commitment has its roots in the way human pro-social behaviour has been evolutionarily shaped.

Chapter 4. Normative Expectations

The aim of this Chapter is to analyse the capacity of social commitments to create normative expectations. When an agent is requested, commanded, or has promised to do φ, the creditor, to whom the commitment is made, normatively expects that the debtor does φ. In fact, I will argue, a social commitment entails an agreement on the claim that "the debtor ought to do φ" (§4.2). But first I will address a closely related problem: the source of the normative expectations created by social commitments. I will analyse three solutions to the so called *problem of promising*, which can be generalised to other kinds of social commitments: how can one put oneself under the obligation of doing something merely by promising? This problem dates back to Hume's *Treatise of Human Nature*[1], and is still a source of debate. Section 4.1 explores three possible solutions to this problem: practice-based, expectation-based and reasons-based accounts. In §4.2, I develop a reasons-based approach to the problem of the normativity of social commitments.

4.1. The source of the normativity of social commitments

Social commitments give rise to normative expectations. When I promise my sister to walk her dog, it is normatively expected from me to walk Godard, her dog: anyone who is aware of the promise can infer that I ought to walk Godard, or, in a weaker conception, that I have a normative reason to walk Godard. However, the fact that social commitments are the source of normative expectations does not entail that social

1 [1739] (2007: 1:)

commitments are normative themselves. Regarding the specific case of promises[2], Vallentyne[3] distinguishes between normativized and non-normativized conceptions. On the normativized conception, keeping one's promise is obligatory in virtue of the act of promising: the promiser validly offers to undertake an obligation, and the promisee accepts. The non-normativized conception claims the opposite: that there is no obligation derived from the act of promising itself—promising would be equivalent to giving one's word about a future conduct. From this perspective, a promiser is obliged to alert the promisee in case she is aware that she will not fulfil her promise, and to compensate the promisee, but does not have a specific obligation to act as promised. Vallentyne illustrates this distinction through a comparison between the concepts of *murder* and *killing*. Whilst murder is defined as wrongful intentional killing, and is therefore wrong by definition, killing is a non-normativized concept, "even if one believes that killing is always wrong, this is not part of the concept of killing"[4].

Accounts of the normativity of social commitments can be categorised in other ways. Pratt [5] distinguishes between voluntarist and non-voluntarist theories of promising. Following the former, promises entail voluntary obligations, insofar the validity of this obligation lies in the intention of the promiser to acquire it, and the intention of the promiser counts as reason for the existence of the obligation. Voluntarist theories can also be conventional or non-conventional, depending on the source of the normative power. On the other hand, non-voluntarist theories (such as Pratt's) claim that the obligation entailed in a promise is external, and not due to the exercise of a normative power, whether conventional or not.

I will now present three approaches to promissory obligations, which can in principle be extrapolated to the broader genre of social commitments. The classification I

2 Although my aim is to analyse the obligations arising from social commitments in general, I shall mainly take examples of promissory obligations.

3 (2006)

4 Ibid., 10

5 (2007)

will follow is based on the one provided by Watson[6]. First, practice-based views, particularly Rawls', claim that promissory obligations derive from a social practice, an institution of promising, and a principle of fairness justified by the benefits of that social practice. Second, expectationalist views, recently leaded by Scanlon, claim that the obligation to keep one's promises is derived from certain principles that regulate the creation of expectations on others. Lastly, reason-based views are both compatible with conventionalist and expectationalist accounts, because they do not entail a theory about the source of promissory obligations, but claim that these obligations have to be understood in relation with the normative reasons they create.

4.1.1. Practice-based views

Social commitments such as contracts, promises and agreements can be understood as individual acts, and also as social practices. Some authors argue that it is not possible to socially commit oneself other than by invoking a preexisting committing practice. Therefore, social practices of promising, for instance, have explanatory priority over the individual promises. Practice-based views are conventionalist: they stress the conventional origin of these social practices, and explain the obligations involved in them by appealing to the benefits of social commitments for group cooperation[7]. Long-term cooperation, as I have argued in §3.1, relies on trust and credibility, enabled by social commitments. Some authors aim to derive the obligation to keep one's promises from the social benefits entailed by this practice.

Rawls[8] is one of the most prominent practice theorists. His analysis of promises is based on his distinction between *summary* and *practice* views of rules[9]. While summary rules guide behaviour given the previously obtained outcomes, practice rules are logically prior to individual cases, insofar they *constitute* the practice. Thus, rules of promising are

6 (2005)
7 Habib (2008)
8 (1999)
9 Rawls (1955)

practice rules: unless the particular act of promising appeals to the constitutive rules, there is no obligation to fulfil one's promises. There are two basic elements in Rawls' theory of promissory obligations. First, there must be a conventional practice, an institution, of promising, with constitutive rules:

> In the case of promising, the basic rule is that governing the use of the words "I promise to do X." It reads roughly as follows: if one says the words "I promise to do X" in the appropriate circumstances, one is to do X, unless certain excusing conditions obtain. This rule we may think of as the rule of promising; it may be taken as representing the practice as a whole. It is not itself a moral principle but a constitutive convention. [10]

For Rawls, constitutive rules, which are public, define institutions. However, it is not from those rules that the obligation to keep one's promises emerges; rather, Rawls invokes a universal principle, the *principle of fairness*:

> [The principle of fairness] holds that a person is required to do his part as defined by the rules of an institution when two conditions are met: first, the institution is just (or fair), that is, it satisfies the two principles of justice; and second, one has voluntarily accepted the benefits of the arrangement or taken advantage of the opportunities it offers to further one's interests.[...] We are not to gain from the cooperative labors of others without doing our fair share. [11]

Rawls' contractualist approach is based on a logically original position, where this principle is agreed because it gathers the benefits of social cooperation. The principle of fidelity is derived from the principle of fairness, and it applies to *bona fide* promises:

> [A] bona fide promise is one which arises in accordance with the rule of promising when the practice it represents is just. Once a person says the words "I promise to do X" in the appropriate circumstances as defined by a just practice, he has made a bona fide promise. Next, the principle of fidelity is the principle that bona fide promises are to be kept. It is essential, as noted above, to distinguish between the rule of promising and the principle of fidelity. The rule is simply a constitutive convention, whereas the principle of fidelity is a moral principle, a consequence of the principle of fairness. […] The obligation to keep a promise is a consequence of the principle of fairness. [12]

Therefore, Rawls' theory of promises is non-normativized, or externalist: there is something wrong in breaking a promise—the violation of the principle of fidelity—but

10 Rawls (1999: 303)
11 Ibid., 96
12 Ibid., 304

this wrong is not intrinsic to the act of promising, which is no more than a conventional social practice.

Two aspects of Rawls' social practice view have been extensively criticised[13]. First, following conventionalism, it is not possible to make a promise in the absence of a practice of promising. However, the origin of this practice can be hardly understood in the absence of any acts of promising[14]. Promises do not exist if there are no constitutive rules[15] that allow to describe that act as a promise:

> Unless such concepts as future, promise, and obligation are grasped by members of a community, at least in a primitive way, there is little chance that rules defining these concepts can be understood by members of the community. But if such rules are understood, then the actions or concepts they purportedly define must exist prior to the rules. The result is that the success of the institutional approach depends on presupposing the concepts and actions that the rules purportedly define. [16]

Second, the directed nature of promissory obligations is ignored in the social practice view. When I promise you to do something, I have an obligation towards *you*; and if I break my promise, I wrong *you*. Promissory obligations are owed to specific individuals, i.e. the persons to whom the promises are made. In the social practice view, the wrong of breaking a promise is, basically, that the agent is free-riding, and is thus directed to the collective relying on this practice. But, as Watson puts it, "If I break a promise, the promisee has a special complaint that goes beyond the accusation that I have exploited or taken advantage of a just institution"[17].

13 see Scanlon (1998: chap. 7)

14 Thomson (1990: chap. 12); see also Sheinman (2011)

15 Despite being widely used, Morin (2009) has recently criticised the concept of constitutive rule, arguing that constitutive rules are historically and socially shaped, and that the apparent social consensus that underlies these rules is often conceals their changing character; these arguments would also apply here, although I will not focus on them.

16 Vitek (1993: 46)

17 Watson (2005: 7); Pratt (2007); see also Scanlon (1998: 316)

4.1.2. Expectation-based views

Scanlon[18] is considered the most prominent defender of the expectation-based account[19]. As opposed to social practice views, expectation-based accounts claim that there is no need of a promising practice for a promising act to exist, or to be normatively binding. However, they do not claim that the obligation entailed in a promise is intrinsic to the promise itself, but it is based on the obligations in which we incur when we lead others to form certain expectations about our conduct. The wrong of breaking a promise would be derived from the wrong of betraying trust, or assurance, which is considered as valuable:

> [T]he wrong of breaking a promise and the wrong of making a lying promise are instances of a more general family of moral wrongs which are concerned not with social practices but rather with what we owe to other people when we have led them to form expectations about our future conduct. [20]

Hence, obligations (and the wrong derived from their violation) are directed towards the agent to whom a promise has been made. The moral obligation generated by a promise is derived from what Scanlon calls the *Principle of Fidelity*, or Principle F:

> Principle F: If (1) in the absence of objectionable constraint, and with adequate understanding (or the ability to acquire such understanding) of his or her situation, A intentionally leads B to expect that A will do X unless B consents to A's not doing so; (2) A knows that B wants to be assured of this; (3) A acts with the aim of providing this assurance, and has good reason to believe that he or she has done so; (4) B knows that A has the beliefs and intentions just described; (5) A intends for B to know this, and knows that B does know it; and (6) B knows that A has this knowledge and intent; then, in the absence of special justification, A must do X unless B consents to X 's not being done. [21]

Following Scanlon's contractualist approach, Principle F would be universally accepted given the value of assurance, which gives rational agents to a reason to accept this principle. Also, there are other ways, besides promising, to incur in an obligation to fulfil

18 (1990); (1998); (2003: chap. 13)

19 However, the idea that expectations give certain rights and obligations is much older; see for instance Sidgwick (1907). I will focus on Scanlon's account, because his theory has been the object of a debate during this last decade.

20 Scanlon (1998: 296).

21 Scanlon (2003: 245). This formulation of the Principle of Fidelity is more recent than the one provided in Scanlon (1998: 304), although it varies little.

Principle F. In general, social commitments create normative bounds derived from this principle.

The problem of this view, as noticed by Scanlon himself[22] and others[23] is that it is subject to a vicious circularity. The problem can be stated as follows. Scanlon claims that the wrong of breaking a promise is due to a violation of a general principle, which applies when we create expectations on others. A promise can create such expectations when the promisee believes that the promiser has a normative reason to keep the promise, insofar Principle F applies. But this principle only applies when those expectations have been created. For the promisee to rationally expect from the promiser that the promise will be kept, the promiser has to communicate her awareness of being under an obligation to do as promised, or at least, that she has a normative reason to do as promised. Thus, promissory obligations depend on expectations; and, in turn, expectations depend on the existence of promissory obligations[24].

Scanlon's solution to the circularity problem consists in clarifying the relation between reasons and expectations. He appeals to a differentiation between two kinds of wrongs: (i) unjustified manipulation, and (ii) the attempt to commit the first kind of wrong. These two wrongs are a violation of two principles: (i) Principle of Unjustified Manipulation (or Principle M) and (ii) Principle of Due Care (or Principle D):

> Principle M: In the absence of special justification, it is not permissible for one person, A, in order to get another person, B, to do some act, X (which A wants B to do and which B is morally free to do or not do but would otherwise not do), to lead B to expect that if he or she does X then A will do Y (which B wants but believes that A will otherwise not do), when in fact A has no intention of doing Y if B does X, and A can reasonably foresee that B will suffer

22 Scanlon (1998: 307).

23 See Kolodny and Wallace (2003); Watson (2005); Mason (2005); Tognazzini (2007); Rivera-Lopez (2006).

24 The circularity problem reflects a more general problem concerning social strategic commitments, which has been stated in Chapter 3. As Sánchez-Cuenca points out, the reason why a commitment is needed in the first place is that the agent lacks of sufficient reasons to do what is promised—otherwise, it would be unnecessary to set up a commitment. However, if the agent lacks sufficient reason to do what she promises to do, then she also has reasons to break the promise. This would also entail a kind of circularity: if an agent needs a commitment device to motivate her to do X, then it is not clear why promising to do X will make the claim that the agent will do X credible. As Scanlon notes, "Typically, a promise is asked for or offered when there is doubt as to whether the promiser will have sufficient motive to do the thing promised. The point of promising is to provide such a motive" Scanlon (1998: 322).

> significant loss if he or she does X and A does not reciprocate by doing Y. [25]

> Principle D: One must exercise due care not to lead others to form reasonable but false expectations about what one will do when one has good reason to believe that they would suffer significant loss as a result of relying on these expectations. [26]

The argument goes as follows. Although promissory obligations arise from Principle F, this principle is not needed to explain the capacity of promises to give the promisee a reason to form the expectation that the promiser will fulfil her promise; Principles M and D can provide the assurance that is needed to trigger Principle F. When I promise my sister to walk her dog, I am both expressing my intention to walk her dog, and also communicating that I take promises seriously—this is, that I do care about not violating Principles M and D. If my sister is aware of this, she will believe that I have a distinctive reason to walk her dog, and this reason would not derive from Principle F. And, from this belief that I have a moral reason to walk her dog, she can reasonably expect that I will do so: it is now that I acquire a promissory obligation towards her.

Scanlon's solution to the circularity problem has been criticised, mainly because of its inadequacy to account for the special kind of assurance provided by a promise. For example, Pratt [27] argues that Scanlon's argument fails in that either Principle F is indeed needed to create the right kind of expectations—and therefore the argument is circular—or it is false. Pratt focuses on two aspects of social commitments: first, the conditions under which an agent who makes her intentions public can change her mind, and the rights conferred to the promisee by the act of promising. Promising requires that the promiser leads the promisee to believe that, unless the promisee consents to the promiser not so acting , the promiser is obliged to do as promised. Principles M and D can lead to forming an expectation, but the speaker is always entitled to change her mind, in which case she ought warn the hearer, and maybe compensate for possible losses, but that is all—

25 Ibid., 298.
26 Ibid., 300.
27 Pratt (2002).

the speaker is not obliged to perform as asserted, which would be a distinctive feature of promises, according to Scanlon.

A similar critical argument has been raised by Kolodny and Wallace[28]. The problem with Scanlon's account, they argue is that the promisee has no reason to believe that the promiser will adhere to her intention. Principles M and D can make the promisee believe that the promiser intends to do what she promises to do, but not that she will not change her mind. Unless the promiser had a compelling non-moral independent reason for doing what she promises, the promisee cannot obtain the *assurance* needed to consider the promiser's assertion a promise—it is necessary that the promisee has a reason to believe that the promiser's intention will persist:

> [T]he distinctive utility of promising is not simply that it allows A [the promiser] to assure B [the promisee] that A will do X when A has prior or nmp [non-moral practice based] reasons to do X that he prefers not to communicate to B, but also that it allows A to assure B when A does not have *any* prior or nmp reasons to do X at all […] On Scanlon's account, promises can no longer perform this vital service. If A can promise to do X only by leading B to believe that A has some prior or nmp reason to do X, then in a situation in which A has no prior or nmp reason to do X, A cannot promise at all. [29].

Therefore, while Kolody and Wallace agree in that the wrong entailed by a violation of one's promissory commitments exceed the wrong of deviating from a social practice, they claim that the obligation to keep one's promises has to be understood in terms of the conventional character of promising. Thus, they suggest a hybrid view between practice and expectation-based accounts[30].

To sum up, according to social practice and expectation-based views, a promise—although their analysis can be extended to other social commitments such as agreements and contracts—is a public declaration of intentions, over which a (moral) principle applies, making the fulfilment of the promise obligatory. This principle can be either conventional, based on the social practice of promising, or an idealization of a moral maxim, which, on Scanlon's contractualist account, is a principle that every rational agent

28 Kolodny and Wallace (2003).

29 Ibid., 143, their emphasis.

30 See Tognazzini (2007) for an alternative hybrid account; Tummolini et al. (forthcoming) also offer a hybrid view, which combines the conventional origin of expectations and a shared value for reliability.

would accept. Also, both the social practice and the expectation-based accounts focus on the role of promises and other social commitments in sustaining and promoting cooperation and assurance. In Chapter 3 I provided an account of the mechanisms that enable social commitments, and concluded that those mechanisms—basically, reputation and emotions—serve to promote trust and cooperation, and are able to explain (i) why people make social commitments, and (ii) why people tend to fulfil those commitments. I fully agree in that the evolutionary origin of the practice of promising, as well as other social bonds, has its roots in the promotion of cooperation. Also, in that this evolutionary origin can shed some light in the moral character of commitments. However, I believe that social commitments have a normative structure which can be analysed independently of their moral character. I will discuss now the third approach to promises, which relates de obligation entailed in promises with their capacity to create reasons for action.

4.1.3. Reasons-based view

The reason-based view is a family of views that defends that promises, as well as other kinds of social commitments, consist in the voluntarily acquired obligation of an agent to perform an action, which is acquired through the exercise of normative powers. Furthermore, by voluntarily putting herself under the obligation to do something, the agent has created a normative reason to do so. In fact, normative power is "the power to effect a normative change. A normative change can be interpreted to comprise every change in the reasons that some person has"[31].

Reason-based views are compatible with different claims about the origin of normative powers. Following Thomson, the mere intention to (morally) bind oneself is enough to have the capacity to bind oneself:

> There is nothing deeper that either needs to be or can be said about how word-givings generally and promisings in particular generate claims. Their moral force lies in their

31 Raz (1975: 99)

> generating claims; and the fact that they do generate claims is explained by the fact that issuing an invitation is offering to bind oneself, so that when the invitation is accepted, the offer is accepted, and one therefore is bound. [32]

Similarly, Searle argues that promises create desire-independent reasons for action, and creating such kind of reasons is available to the agent as an exercise of her will:

> The obligation to keep a promise does not derive from the institution of promising. When I make a promise, the institution of promising is just the vehicle the tool that I use to create a reason. The obligation to keep a promise derives from the fact that in promising I freely and voluntarily create a reason for myself. [33]

The problem with these accounts is the capacity to create normative reasons at will, even in the absence of normative reasons to do so, or in the presence of contrary normative reasons. First, if social commitments provide reasons, these reasons are somehow disconnected from the content of the commitment, because they do not rely on its desirability or goodness: requests, commands and promises give rise to content-independent reasons for doing what is requested, commanded, or promised:

> Content-independence of commands lies in the fact that a commander may issue many different commands to the same or to different people and the actions commanded may have nothing in common, yet in the case of all of them the commander intends his expressions of intention to be taken as a reason for doing them. It is therefore intended to function as a reason independently of the nature or character of the actions to be done. [34][35]

To put it otherwise: being requested to do something creates a reason to do it, independently of the content of the request. This characterization of the reasons created by social commitments leads to accepting that, upon request, there would be a normative reason for anything that could be requested, including prohibited or immoral actions. This seems to deflate the justificatory power of normative reasons. Closely related, it can be argued that some social commitment, such as immoral promises, are not binding, because there cannot be a good normative reason to do something that one ought not to

32 Thomson (1990: 303)

33 Searle (2001: 198)

34 Hart (1982: 254).

35 Quoted in Gur (2011). See also Raz (2001) for a characterization of content-independent reasons and justifications.

do, everything considered. Gilbert[36] takes this claim as a dogma of the theories of promises, and argues, against it, that immoral promises are *non-morally* binding, so they give rise to obligations, but obligations of the *moral* kind. Finally, Watson[37] argues that normative powers have limits. An agent is not obliged to perform an action she was not entitled to do in the first place. The *Autonomy Thesis* is a limitless version of the reason-based account of promissory obligations, stating that "that we have the power to make valid promissory commitments to anything whatever"[38]. The capacity to create normative reasons without limits seems to be a case of bootstrapping (see §2.1.3): if there are no normative reasons to φ, just by intending I do not create a new reason to do so. However, what happens in the case of requests or commands?

In the following Section, I will defend an account of the normativity of social commitments which broadly belongs to the reason-based view. I believe that promises, requests, or commands, create normative reasons for action that did not exist before the commitment was made. I will try to clarify what kind of reasons they create, and why they are normatively—although not necessarily morally—binding. I believe that the moral status of certain promises relies on social values such as friendship, reliance, and trust, and that these values have a conventional origin. As I explained in Chapter 3, no doubt social commitments enhance cooperation, and trust and reputation make social commitments credible and effective. However, I do not want to commit myself to the view that a social commitment creates a moral obligation because it serves morally good values, such as reliance. Of course, if social commitments did not give rise to empirical expectations, this is, if no one believed that the promiser is going to fulfil her promise, they would not give rise to normative expectations either. Furthermore, it is quite plausible that the explanation of how normative powers are possessed is conventional and practice-based. My claim is that social commitments, as well as individual commitments, have a normative dimension which is independent of their moral dimension, in the sense that it

36 Gilbert (2011).
37 Watson (2009).
38 Ibid., 169.

concerns a relation amongst intentions, reasons, judgements and actions that is susceptible of being correct or incorrect—not morally, but rationally. This is not to say that rationality requirements do not have a socially shaped form. After all, they are the tool with which we interpret and make sense of our actions, and other agent's actions. Thus: even if the content of out moral values, the fact that we we have moral and social norms that regulate our social (and individual) life is evolutionarily shaped, as well as our rationality, it does not mean that social commitment do not have a normative structure worth exploring. My aim is to analyse that structure, and to relate it to the rationality requirements governing practical commitments.

4.2. Reasons, obligations and entitlements

When I promise my sister to walk her dog, she does not only empirically expect that I will do it: she also believes that I ought to do it, and moreover, she believes that I ought to do it *because* I promised her to do so. We may have normative expectations over an agent because we believe that this agent has good normative reasons to do something, and also because we believe that the agent ought to do something because she owes it to us. In this latter case, we stand in a particular normative relationship with that agent. These two kinds of normative expectations arise from different kinds of commitments of the agent: propositional and agential commitments. If my sister tells me that she wants to improve her English, because that will afford her an opportunity to apply for better jobs, and therefore she is going to sign up for an English course, she is acquiring an assertoric commitment. I believe that, all things considered, she ought to sign up for that course: I normatively expect she does so. But if she does not, she is not violating any commitment she had with me: she can be asked for an explanation on why she has changed her mind, but she is not requires to sign up to the course *because* she has communicated her intention to do so. To put it otherwise: the fact that she told me about her intentions does not constitute a reason for her to do what she told she was going to do. On the other

hand, having promised to do something does constitute a reason for doing what is promised. The following pages are devoted to analysing why this is so.

4.2.1. Propositional and action commitments

> [Wle are not bound to make our actions correspond with our assertions generally, but only with our promises. If I merely assert my intention of abstaining from alcohol for a year, and then after a week take some, I am (at worst) ridiculed as inconsistent. [39]

My aim in this §is to draw a distinction between two kinds of public commitment: *propositional* and *action* commitment[40], also called *doxastic* and *practical* commitment[41]. It is widely accepted that promissory obligations differ from those obligations arising from assertions. However, some authors, specially those who endorse what Watson[42] calls the *Autonomy Thesis*, (see §2.1.3), have difficulties in explaining the difference. From this perspective, promises are analysed in terms of the undertaken obligations that follow from specific speech acts. The Autonomy Thesis claims that an agent can freely create a normative reason for action by voluntarily undertaking an obligation; for example, by promising:

> When I say "I promise to wake you at 6:00 A.M.," I see myself as freely creating a special type of desire-independent reason, an obligation, for me to wake you at 6:00 A.M. This is the whole point of promising. Indeed, that is what a promise is. It is the intentional creation of certain sort of obligation—and such obligations are by definition independent of the subsequent desires of the agent. [43]

> Consider a person's making a promise by saying "I promise to bring the book back tomorrow" or giving a name to a ship by saying, in appropriate circumstances, "I hereby give you the name 'Nautilus'". This person is doing things by words, in a sense. A new state of affairs involving obligations and rights is created by these utterances, a state involving, respectively, a commitment to keep the promise and a collective commitment to use the given name. Generally speaking, all this happens by virtue of some accepted principles,

39 Sidgwick (1907: 304).
40 Walton and Krabbe (1995).
41 This terminology is from Brandom (1994).
42 Watson (2009).
43 Searle (2001: 209).

indeed collectively accepted principles. [44]

From this perspective, promising would be a special kind of assertion, one that creates new reasons for action. I agree with this so far. However, while an assertion commits the agent to justify what is asserted, a promise commits the agent to do what is promised, and this requires the acceptance of the promisee. The role of the creditor is often underestimated in the literature on social commitments. The traditional view on promising is that it is a communicative act that declares a commitment to a future intention, and agents can perform it autonomously[45]. The addressee does not play a role in this analysis. However, action commitments require the hearer's acceptance[46]. Suppose that Searle promises me to wake me at 6:00 A.M.; however, I (politely) reject his offer, because I have no reason to wake up at 6—moreover, I want to sleep until 8:00 A.M. Is Searle obliged to wake me at 6 just because he said so? It seems not; he has not acquired an obligation towards me. Promises are, in this sense, directed obligations: the promiser has an obligation towards the promisee, and not towards anyone else. Witnesses may be entitled to require justification, insofar propositional commitments are not directed (although of course they are in principle restricted to those that hear, or read, the assertion).

In order to examine the differences between propositional and action commitment, I will use four examples. The first two cases are examples of propositional commitment, involving a declaration of intentions; the third, however, seems to incur in a wrong by not doing what is asserted. The fourth example is an explicit action commitment.

44 Tuomela and Balzer (1998: 177).
45 Clark (1996: 125ff.).
46 Bouvier (2007).

Example 1: Asserting	**Example 2: Unintended expectations**
Ann: "What are you doing next summer?" Bob: "I'm going one week to Lanzarote with my family" Next summer, Bob and his family change their plans and go to London for a week instead.	Ann: "Will you use your car tomorrow?" Bob: "No, I'll walk to the office." The day after, it is raining and Bob decides to take his car; unbeknownst to him, Ann intended to take it as well, because her car broke down. Ann has to phone a friend to get picked up and driven to the airport.
Example 3: Reliance	**Example 4: Agreements**
Ann: "Will you use your car tomorrow?" Bob: "No, I'll walk to the office." The day after, it is raining and Bob decides to take his car; however, he knew (because he heard Ann talking on the phone with a friend) that Ann was planning to get to the airport with his car. Ann knows that Bob was listening to her conversation. Ann has to phone a friend to get picked up and driven to the airport.	Ann: "Would you lend me your car tomorrow? Bob: "Sure, I'll go walking to the office" The day after, it is raining and Bob decides to take his car; Ann has to phone a friend in order to get a ride to the airport.

Table 5: Assertoric and action commitments

Assertoric commitments and justificatory responsibility

Example 1 is a clear case of public declaration of intentions. Making an assertion is an action (a speech act) consisting in presenting something as what is the case, stating a fact. Of course, a fact can refer to a past, present or future state of the world. Public declarations of intentions refer to future events that the speaker intends to achieve.

Asserting commits the speaker to what is asserted; however, there are different ways to understand the relation between agent and object in this kind of commitment. One possible way of understanding this commitment is to relate it to the responsibility taken by the agent. For example, following Searle, the commitment in which an agent incurs when asserting a proposition consists in taking responsibility for three facts:

> In making an assertion we take responsibility for truth, sincerity, and evidence [...] These responsibilities are met only if the world is such that the utterance is true, the speaker is

sincere, and the speaker has evidence for the assertion. [47]

This notion of responsibility is quite obscure, as Rescorla[48] points out. First, it does not distinguish between prospective and retrospective responsibility. While the truth of a public declaration of intentions depends on whether those intentions are translated into action, and it is therefore a kind of prospective responsibility, having been sincere or having had evidence for what is asserted is a kind of retrospective responsibility. Once I have asserted that p, I have done so either believing that it was true or false, or lacking of enough evidence for my assertion. Thus, I can be responsible for lying, or for presenting something as true while not having enough evidence. Searle identifies, or at least closely relates, responsibility (in the prospective sense), commitment and obligation. In the case of declarations of intentions, as in Example 1, the felicity conditions for what is asserted are not clear. It would seem that the agent needs to perform what is asserted in order to fulfil his commitment. Indeed, it is a very strong condition: changing one's mind entails violating the agent's commitment to her assertion.

On the other hand, some philosophers have proposed that the kind of commitment involved in the act of asserting is a commitment to defend, argue for, or justify what is asserted [49]. Similarly, for Walton and Krabbe "to assert a proposition may amount to becoming committed to subsequently defending the proposition, if one is challenged to do so by another speaker"[50]. The responsibility the agent takes with respect to what she asserts (p) does not require that the agent makes p to be the case. Watson argues that, by means of a promise, the agent is committed to making it true what is promised. In the case of asserting, the agent is committed to the truth of p. Along the same lines, Carson claims that "[t]o warrant the truth of a statement x is not necessarily to place oneself under an obligation to make it true that x"[51].

47 Searle (2001: 176)
48 Rescorla (2009).
49 See Brandom (1983); (1994); Watson (2004b).
50 Walton and Krabbe (1995: 31).
51 Carson (2006: 294).

Hence, there is a difference between the kind of commitment created by means of an assertion, and the kind of commitment in which we incur when making a promise, for example. Walton and Krabbe label the former *action commitment*, and the latter *propositional commitment*. In Brandom's terms, an action commitment is a *practical* commitment: it is a commitment to intend and to act under the light of the reasons one has. A propositional commitment, on the other hand, is a *doxastic* commitment:

> [A]ssertings [...] are in the fundamental case what reasons are asked for, and what giving a reason always consists in. The kind of commitment that a claim of the assertional sort is an expression of is something that can stand in need of (and so be liable to the demand for) a reason; and it is something that can be offered as a reason […] Other things besides assertional commitments involve liability to demands for justification or other demonstration of entitlement—for instance, the practical commitments involved in actions. [52]

Example 1 above contains an assertion that involves a future state of affairs that is intended by the agent. Bob asserts that he will go to Lanzarote the following summer. What happens if he does not? Is he violating his commitment? My suggestion is that commitments to the assertion that p are not violated if they are either fulfilled or cancelled. They can be fulfilled by justifying p, or cancelled by retracting p and justifying not-p; thus, changing one's mind does not violate this commitment.

Commitments can change their normative state. Not fulfilling a commitment does not necessarily entail that the commitment has been violated. For example, if certain conditions are met, commitments can be cancelled. In Example 1, Bob can freely change his mind; Ann is entitled (in Brandom's sense) to ask Bob, when the time comes, why he has not gone to Lanzarote. Bob can simply answer that he thought that going to London would be a better idea, and so he and his family changed their plans. In any case, asserting that he will go to Lanzarote commits Bob to explaining why he has or has not done that when the time arrived. It is a propositional commitment: once Bob has asserted his intentions, the hearer in entitled to ask for justification of what is asserted, whether what has been asserted corresponds to an actual state of the world or not.

52 Brandom (1994: 167).

Example 2 is a bit more complicated, for it entails expectations and reliance. As Brandom argues, a speaker's assertion entitles the hearer to make inferences taking what has been asserted as reasons. In the described scenario, Ann has inferred that the car would be available the day after. She has relied on Bob's statement, and suffered from a loss as a consequence of a violation of her expectations. However, Bob did not know that Ann was relying on his words in order to get to the airport: he did not know that Ann's car war broken, neither that Ann had to take a flight the day after. For the sake of simplicity, I will assume that, besides not actually knowing about her intentions, it was not reasonable for Bob to know about her intentions. The day after, Ann may ask Bob why he finally took the car, having said that he would not. Bob does not violate his propositional commitment if he explains that it was raining, so he changed his mind. Bob wanted to walk to the office because he wanted to enjoy a walk; the fact that it is raining is a reason to prefer taking the car instead, for the previous reasons for going walking are not met any more. Ann can be upset because the car was not available, but she is not entitled to make Bob responsible (in the sense of accountable) for her loss. She could have let Bob know that she was planning to take the car. Bob has unintentionally created expectations on Ann. In Scanlon's account, Bob would not be violating Principles F, M or D, because he is not intentionally misleading Ann concerning the use of the car. In fact, expectation-based accounts stress the fact that agent who creates the expectations has to be, at least, aware of her having created them.

In a nutshell, there are two important features of propositional commitments shown in the examples. First, cancelling a commitment is different from fulfilling it, and in both cases, it is a matter of the relation between the speaker and her actions. Making a public declaration of intentions does not bind the agent to her intentions more than just having that intention. This is, once the reasons for performing what one has asserted change, or are not into play any more, there is no further reason to perform it. Asserting is a form of reason-giving, but does not create a reason that explains or justifies the content of the assertion. An agent can legitimately (rationally speaking) change her mind, thus cancelling her commitment (i.e. her intention), without incurring in a rational mistake.

On the other hand, the commitment a speaker has towards the hearer is not violated either, as long as the agent justifies her assertion, providing the reasons to intend to do something in the future, and justifies (in the case that the action is not performed) the mismatch between what is asserted and her actual actions.

The transition from a propositional commitment to an action commitment is not an all-or-nothing matter. There is a wide scope between trivial declarations of intentions to strong, binding promises. The difference lies on the expectations created, allowed and offered. Sometimes, these expectations are created intentionally: that is, in general, the reason why people of make promises. Or, it is possible to assert an intention, stressing that the speaker is not committing herself to its performance ("I believe I will do p, but I don't promise anything"). Between these two cases, there is a wide scope of assertions about future intentions that, more or less implicitly, allow for the creation of empirical and normative expectations on the hearer. As I will argue in the next Section, action commitments involve the hearer's acceptance, and an implicit agreement about the normative expectations concerning the debtor's actions.

Actions commitments and task-responsibility

Example 3 is very similar to Example 2; however, it seems that Bob should have informed Ann that he might take the car, or should have lend his car to Ann. He is, apparently, under some kind of obligation towards Ann, and this obligation is derived both from his assertion (expressing that he would not take his car the day after) and his knowledge about Ann's reliance on him for having the car available to get to the airport. However, what difference does it make that knowledge (for the assertion was also made in Example 2)?

In fact, in Example 3, Bob has a reason to fulfil his intention that did not exist before letting Ann know that he was not going to take the car the day after. Because of this, Bob cannot justify that he took the car the day after simply because he changed his mind. The common knowledge about Ann's intentions and reliance makes of this case an

example of tacit agreement [53]. Ann acquires some authority over Bob's use of the car [54]. Bob's responsibility is not limited to justify what he has asserted, but extends to making his assertion true.

Example 4 is a standard case of explicit agreement: Bob has agreed to lend his car to Ann. It is unidirectional, in the sense that no counterpart is required from Ann.

In the scenarios described in Examples 3 and 4, Bob's commitment will be violated in the moment he takes his car instead of going walking to work. Even if he has very good normative reasons that make it reasonable and justified to change his mind, the social commitment created between Ann and Bob is not cancelled nor fulfilled. It might even be the case that Bob should, all things considered, take his car to get to work; it is also possible to make promises to do things we should not do. The point is that, while the propositional commitment is not violated by changing one's mind, the action commitment is. Ann has the right to require the use of the car, and to demand good reasons for breaking his commitment to Bob.

To summarize, the difference between propositional commitment expressing future intentions and action commitments to other agents relies on what is the object of the commitment. Propositional commitments are directed towards the justification of what is asserted, while the object of action commitments are precisely the performance of certain action: "In asserting my future intentions, I express my mind; in a promise, I commit my mind"[55]. While action commitments create normative reasons for action, and thus play a role in practical reasoning, propositional commitments require to provide the reasons supporting what is asserted when requested by the hearer. While propositional commitments entail a justificatory responsibility, action commitments involve to take responsibility for performing the action that is the object of the commitment. This is, the debtor acquires an obligation towards the creditor regarding the fulfilment of the commitment. A social commitment to perform certain action, thus, is a kind of

53 Tummolini et al. (forthcoming) forthcoming
54 Owens (2006)
55 Watson (2004b: 63).

prospective responsibility. As I will argue in §5.3, bearing this kind of responsibility over a future outcome entails to hold a specific relation with the outcome in which attributions of retrospective responsibility—this is, responsibility for something that already happened—are founded. Now, I turn to analyse the normative elements involved in a social commitment; hereinafter this concept will be exclusively used to refer to action commitments.

4.2.2. The normative structure of social commitments

Promises, requests, commands, agreements and contracts, amongst others, share a common normative structure. Social commitments entail a specific relation between reasons, rational authority, and normative requirements. I will adopt here Castelfranchi's approach to social commitments, which he calls *S-Commitments*. My aim is to connect their normative structure with the normative structure of individual practical commitments. Following Castelfranchi[56], if X (the debtor) is socially committed to Y (the creditor) to do φ, then:

(i) X adopts φ, which is a goal of Y, with her acceptance, which can be explicit or implicit.

(ii) *(i)* is a normative reason to φ.

(iii) A set of rights and obligations is created along with the commitment.

I will first address to the adoption of goals entailed by a social commitment. Then, I will analyse the obligations and entitlements emerging from the commitment, and argue that they are entrenched in the structure of the normative requirements of practical commitments. This is so because social commitments provide normative reasons for performing the action to which the debtor is committed, and the debtor confers her rational authority over these reasons to the creditor.

56 Castelfranchi (1995).

Goal adoption: the importance of uptake

Social commitments are a form of goal adoption[57]. Goals are the content of social commitments, the action to be performed. When an agent publicly states her intentions, she is making her goals public to other agents. However, promises and contracts involve not only the publicity of the goals in question, but also two distinct features of this kind of social interaction: goal adoption and goal delegation. Goal-adoption consists in forming the intention of pursuing the other agent's goal. Goal-delegation, on the other hand, consists in the disposition to allow the other agent to pursue a goal on our behalf. The role of the creditor in the adoption of goals is central to this kind of commitments. The creditor must accept that the debtor endorses that goal: "a crucial condition of that analysis is that the recipient indeed wants the assurance that the agent will perform as indicated, and that the commitment is made for this reason. This is not in general true of the promissory commitments involved in assertion"[58]. For example, if I request you to read a paper I have written, and you accept, you are adopting a goal I have (that you read my paper). The same goes with commands, and also with promises. Thus, goal-delegation determines whether the promise constitute an agreement of goal-adoption. If the promisee does not want that the promiser adopts her goal, or her agreement is irrelevant, then it is not a promise but a mere declaration of intentions. Thus, goal-adoption is an action that occurs at the social level, not at the individual one (although sincere promises encompass both). The sharing of a goal, as Castelfranchi[59] points out, is a basic pro-social structure that is bilateral, although not symmetrical: each agent stands in a different relation to the goal.

Thus, not every instance of goal-adoption is necessarily a social commitment. Goal-adoption can be done privately, and unbeknownst to the agent whose goal is adopted. An agent can intend to perform an action that she believes the other agent intends, without consent of that agent. This is why goal-delegation is important to

57 Ibid.; see also Castelfranchi (1998); Castelfranchi and Guerini (2007).
58 Watson (2004b: 67).
59 Castelfranchi (2008).

constitute a social commitment: "[w]ithout such (often implicit) agreement (which is a reciprocal S-Commitment) no true S-Commitment of x to y has been established"[60]. On the other hand, goal-delegation has also to be agreed. Example 2 illustrates a failure to agree that there is a goal that has been delegated, and adopted. Ann has the goal of borrowing Bob's car, but Bob is not aware of this fact: he has not agreed, neither explicitly nor implicitly, to lend his car to Ann.

The agreement entailed by a social commitment has two features: first, it can be implicit or explicit; second, it can be honest or dishonest.

In example 3, Bob and Ann share certain beliefs that allows for settling an implicit agreement. In a recent article, Tummolini et al.[61] have suggested that the following conditions have to be met for an implicit agreement to take place. First, it is *commonly known* that Ann has decided to *rely* on Bob. Bob has not *disconfirmed* her belief (he omits further information that might lead Ann to changing her belief). The *salience* of Bob's *silence* justifies a common belief in that the expectations on his behaviour are confirmed[62]. Lastly, Bob is not only confirming Ann's expectations, but he is implicitly giving consent to Ann to take his car: "one's consent (not just a behaviour that happens to consent) to the fulfilment of a goal of another agent amounts to the intention not to interfere with another agent's goal fulfilment since and until the other has such goal"[63]. The possibility of creating a social commitment by implicit means can pose some problems, insofar a theory of tacit agreements is needed; but I believe that the approach just presented can provide a general picture of the elements involved. Example 4, on the other hand, shows an explicit uptake by Bob. The above conditions (common knowledge, saliency of Bob's utterance to

60 Castelfranchi (1995: 43).

61 Tummolini et al. (forthcoming)(forthcoming)

62 The authors argue that this step is founded in the nature of conventions: "[E]ach time two or more agents interact with each other in a situation that is governed by a convention, if they keep silent about the expectation of reciprocal reliance that they mutually know to have, each of them tacitly confirms such expectations about each other, even if they are not grounded in direct experience (e.g. the agents might have never met before)" Ibid., 26 forthcoming.

63 Ibid., 29 forthcoming.

the confirmation of Ann's expectations) apply here as well—the only difference is that Bob offers explicit confirmation.

The second remark is that, in the case of promises and other voluntarily acquired social obligations this goal-adoption does not need to be sincere. Otherwise, only sincere promises would be real promises; the agent is supposed to form the intention to achieve the goal, although it is not a necessary condition for assessing whether a social commitment exists. In this sense, social commitments prescribe, but do not necessarily entail, mental attitudes[64]. A dishonest social commitment has the same normative structure than an honest one. Thus, we can infer that individual (practical) commitment is not needed in order to engage in a social commitment. If the agent is being honest, then she intends to do what she is committed to; otherwise, the commitment is still valid, although the agent is being dishonest. An internal practical commitment is neither sufficient nor necessary for a social commitment; on the contrary, the creditor's belief that the debtor is individually committed (i.e. she intends) to do what she is committed to [65].

These two remarks lead to a further step in the analysis. On the one hand, the creditor's uptake, along with the debtor's acceptance to commit herself to do φ, is an agreement, either tacit or explicit. But, what do the agents agree upon? This connects with the second remark just above. The debtor does not necessarily intend to do φ, although she is *supposed* to. And this is precisely what is agreed by the agents: that the debtor adopts a goal (to φ) through the commitment, and that this commitment is, at the same time, a reason for doing φ—the debtor ought to φ because she is committed to the creditor to do so. To put it differently: that having made a promise is a normative reason to do what is promised, or requested.

I will now turn to the problem of the kind of reason that is created in a social commitment, and argue that the normative structure of social commitments reflects the relation between each of the agents involved and the reason created.

64 Castelfranchi (1999).

65 Castelfranchi (1995: 45).

Entitlements, obligations, and normative requirements of rationality

My aim now is to explain the normative relations holding in social commitments, especially those related to the rationality requirements analysed in §2.2. In order to illustrate the problem of promissory obligations, let's imagine the following situation. Bob has promised Ann that he would lend his car to her the day after. Bob and Ann agree in that, because he has promised to, Bob ought to lend his car to Ann when the time comes. Later on, Bob is offered to do a job interview the day after, for which there are many applicants, and he really wants to get the job. He needs the car in order to get to the appointment. However, he does not want to *break* his promise. Is there any way not to do what has been promised, without breaking the promise?

Social commitments, as well as individual ones, have different states. Once created, a commitment can be discharged, cancelled or released[66]. To discharge a commitment is to fulfil it: the debtor does what she is committed to. On the contrary, the debtor cancels the commitment when she either asserts that she is not going to perform what she was committed to, or she simply does not do it. Finally, the creditor can let the debtor off the hook, revoking the commitment, and thus releasing the debtor from any obligation related to that commitment.

In fact, these states are very similar to the possible paths of choice and action after having judged that one ought to do something. The enkratic requirement states that an agent ought not to judge that she ought to do something and, at the same time, to intend not to do it. And the resolve requirement demands from an agent who intends that she persists in her intentions. In both cases, the agent is allowed to *exit from* the requirement, by changing her judgement, or by changing her intention (which is also subject to the enkratic requirement). To exit from the requirement is to revoke the judgement, or the intention, that makes it the case that the requirement applies (see Figures 1 and 2). However, in order to exit from a requirement, one needs to have good reasons to do so. As

66 This classification is based on Singh (1999).

I explained above, an agent can re-evaluate her reasons in the light of new facts that affect the validity of the judgement, or can reconsider the validity of the reason that led the agent to form the judgement in the first place. So, this is how practical (individual) commitments are revoked. Concerning social commitments, only the creditor has the authority to revoke them. My aim is to explain why this is so.

Suppose that there is an person, Merciful Merle, to whom you have committed yourself. Particularly, you have accepted his request to do φ. After the commitment has taken place, you find out that there is a fact, α, that is a reason not to do φ. If you communicate Merciful Merle that you judge that, given α, you ought not to do φ, he will release you from your commitment: you are no longer required to do φ.

Committing yourself to Merciful Merle is pretty much like having a practical, individual commitment. If you believe that there is a reason that cancels your previous judgement in which you base your intention, then you are no longer required, from the point of view of rationality requirements, to stick to your intention: you have exited from the requirement. Letting Merle know this reason will automatically lead him to revoke the commitment.

Let's bring back Bob and Ann's example. Bob has promised Ann to lend his car to her, but he has been appointed for a job interview that he does not want to miss, and he needs his car to get to the appointment. Bob can communicate this fact to Ann, in order to get released. If Ann were Merciful Merle, Bob would be released and he would have not broken his promise, this is, he would not cancel the commitment, but it would have been revoked by Merciful Merle. However, Ann can either judge that Bob's reason is a good reason, or can judge that, in the light of this reason, Bob still ought to lend his car to her. If Ann decides that Bob's interview does not justify changing her judgement, then Bob has no other option than breaking the promise, by unilaterally cancelling the commitment[67]. In any case, it is Ann's decision, because she *has acquired the authority to*

67 A different question would be whether Ann is right in her evaluation of reasons, this is, whether she is right holding the judgement that Bob ought, all things considered, lend his car to her. I will turn later to the subjective / objective distinction on reasons; however, it does not matter for Ann's authority how good Bob's reasons are (either objectively or subjectively). She can be a despotic friend who does not care about Bob's

reconsider the judgement "Bob ought to lend his car to me because he promised to do so", and she has acquired it from Bob. This is the basis for the creditor's rights over the debtor's actions: the authority over a normative judgement.

To sum up what has been said so far: in a social commitment, there is a debtor and a creditor, who agree that the debtor ought to, and ought to intend to, do what she is committed to *because* of having committed herself. With this agreement, the debtor loses her authority to reconsider and re-evaluate the reasons that make it the case that she ought to do what she is committed to; the creditor acquires this authority. Therefore, social commitments create new normative reasons for action that did not exist before the commitment was made but, as opposed to choices, they are not subject to bootstrapping. As Raz pointed out, there are close similarities between decisions and promises:

> That a person promised to do A is a reason for him to do so. One should make a promise only if there are sufficient reasons to do so. But once a promise is made it is a reason for action even though it is a promise which should not have been made. Moreover, a person can promise knowing that he should not. *Once the promise is made he has a reason to perform the promised act* despite the fact that he made the promise knowing that he should not make it. The same is true of decisions. That a person has made a decision is an *exclusionary reason* for him not to consider further reasons. [...] A promise is a reason which can be defeated by other reasons and the fact that it should not have been made may be relevant to whether or not it is defeated. This is true also of decisions. Some will think that a promise is a reason only in virtue of a general principle that *promises ought to be kept*. We could similarly regard decisions as exclusionary reasons in virtue of *a general principle that decisions ought to be respected*. [68]

Here, Raz is discussing the formal similarities between promises and decisions; materially, he argues, they differ. Promises are designed to enable trust and interpersonal predictability, while decisions are designed to settle matters in their own mind and stop deliberating. Furthermore, decisions cannot create new normative reasons for action, as this would be a case of bootstrapping. The difference lies in that, when deciding, we consider and evaluate the reasons we have for and against doing something. Our decisions, as argued in §2.2, are subject to the normative requirements of rationality. Rationality

---

interview: in this case, Bob would be breaking a promise, which might be an unfair promise, but breaking it nonetheless, even for the right reasons.

68 Raz (1975: 69), my italics.

requires that we do not act against our normative judgements, for instance (the enkratic requirement). But rationality does not provide a further reason for not acting against our normative judgements, insofar we have the rational authority to change the judgement[69]. Social commitments, on the other hand, are social facts that, once created, escape the rational authority of the debtor. Once a request has been accepted, for instance, the debtor might think that this was a request that she should never have accepted, and that there are in fact many reasons not to fulfil the request. However, this does not "erase" the request. I can break a promise for very good and strong reasons, and that will not make the promise disappear: I will still have broken the promise.

However, there are two similarities between decisions and social commitments that I want to stress. First, that "once the promise is made he has a reason to perform the promised act", despite any reasons against performing it. This is so because the commitment is a socially acknowledged reason for action, from which a social obligation is derived[70]. The reasons created in a social commitment are in a sense content-independent (see §4.1): having promised to do something is a reason for doing what is promised, independently of the content of the promise. They are reasons insofar they justify the claim that *the debtor ought to do what is promised because of having made a promise*, but they do not directly justify doing that very same action out of the context of a social commitment. In this sense, the reasons created by social commitments are exclusionary. For Raz[71], an exclusionary reason excludes other reasons in the deliberation process leading to a normative judgement about what ought to be done[72]. It is a second-order reason, that does not justify or counts in favour of the action itself, but affects the other reasons the agent may have to perform the action.

---

69 Again, this is not to say that judgement formation and change is not subject to certain epistemic constraints regarding what evidence is acceptable, or the requirement of not to hold contradictory beliefs, for instance. My claim is that an agent is always *entitled* to revise her judgement (even if it is only to find out, again and again, that her judgement remains the same).

70 Miller (2006).

71 Raz (1975); see also Vitek (1993: 89).

72 See Piller (2005) for a critical discussion of exclusionary reasons. I agree with his analysis in that exclusionary reasons may not be as different from attitude-related reasons, although I believe that the distinction is useful, especially regarding socially created reasons.

The main feature of the reason created through social commitments is that it is accepted by the debtor and the creditor (and so the normative judgement is also accepted by both), but only the creditor has the rational authority to reconsider the normative judgement it justifies. Also, a social commitment is a agreement about what counts as an objective reason. The distinction between objective and subjective reasons, as I argued in §3.1 lies in that the latter is a fact that an agent uses to justify or explain her own actions, while the former corresponds to an inference rule regarding what constitutes a reason for or against an action. In the case of a social commitment, the debtor and the creditor agree in that the act of commitment is a reason for performing the content of the commitment. It is expected, thus that the debtor acknowledges that objective reason, and takes the fact of having committed herself to do φ as a reason to do φ.

Regarding Raz's mention of a *general principle* governing both decisions and promises, I believe that normative requirements play here a useful role[73]. In fact, what I am defending is that, if we aim to suggest a general principle of promising (or requesting, commanding, agreeing) it would be a variation of the *resolve* normative requirement. It can be formulated as follows:

> SOCIAL COMMITMENT (NARROW +): If you are socially committed to φ, then you ought to φ in virtue of that social commitment, and you are rationally required to intentionally φ.
>
> SOCIAL COMMITMENT (WIDE -): Rationality requires that you do not [hold a social commitment to φ and intentionally do not-φ]

Thus, the social commitment requirement is a version of the resolve requirement, in which there is a previous judgement about what the agent ought to do (Figure 3 illustrates this case, applied to practical commitment). The agent is thus subject to the enkratic requirement, which forbids to hold intentions that contradict practical normative judgements. The main difference between the resolve requirement and the social commitment requirement is the following: the agent cannot exit from the commitment

73 I diverge here from Raz's approach to the nature of those general principles.

on her own—she has to be *released*. The following figure illustrates the social commitment requirement:

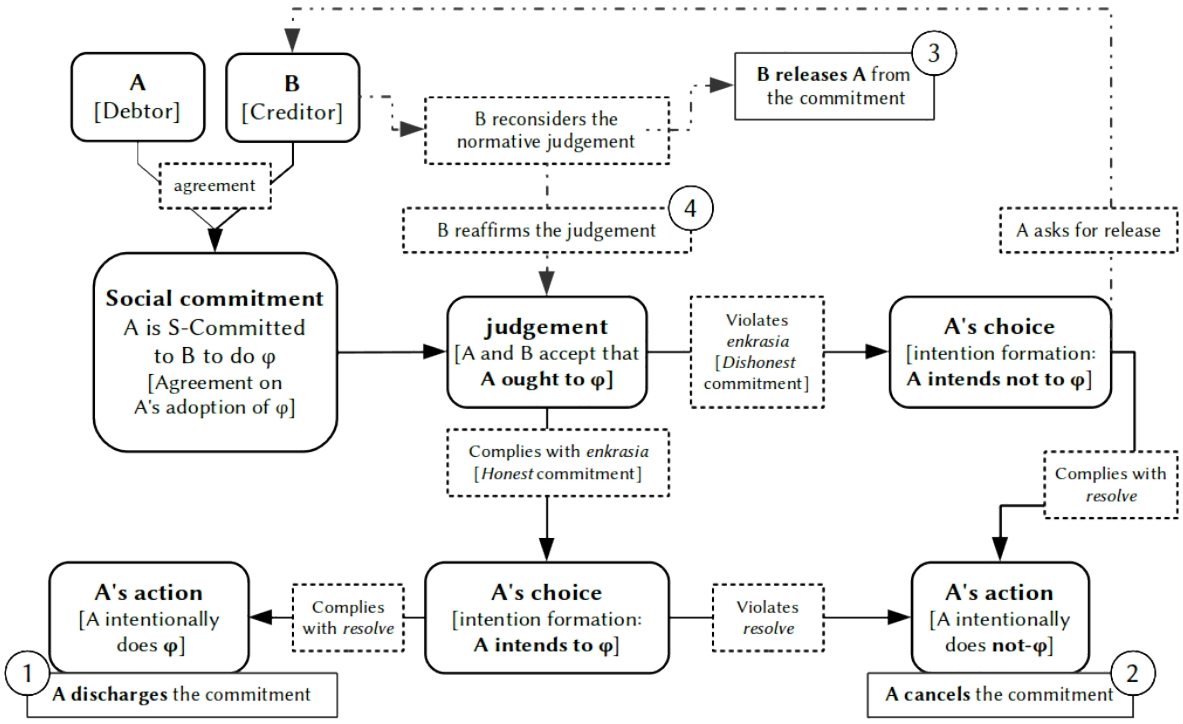

Figure 3: The social commitment requirement

The dotted lines represent the creditor's rational authority over the process. As it is shown, this figure is very similar to Figure 3. The difference lies in that the debtor has lost her rational authority to exit from the requirement. Once a social commitment is created, the debtor can fulfil it (option 1 in the Figure above), violate it (option 2), or ask for being released. The creditor reconsiders then the reasons for achieving the adopted goal, and can release the debtor (option 3) or reaffirm the commitment (option 4). In the case of Merciful Merle, he will release the debtor when she presents a reason she believes to be a good reason; this is why committing oneself to Merciful Merle is identical (rationally speaking) to holding a practical commitment. And this is also why I am sceptical about the possibility of making promises to the self, as a kind of self-commitment that is different from a practical commitment in the form explained in Part I of this dissertation.

The fact that, through a social commitment, an agent delegates her rational authority concerning the obligation to perform what that agent is committed to also gathers the directed nature of the obligations arising from a social commitment. As some authors have pointed out, promissory obligations are directed[74], insofar the creditor has acquired certain rights because the debtor has bestowed those rights on the creditor, as an exercise of her normative powers. As Watson puts is, "the wrong incurred in breaking a promise is the same as the wrong involved in my refusing to relinquish claims to an item I have given you"[75]. A debtor allows the creditor to hold the justificatory authority over her actions (concerning the commitment).

It is also possible that the debtor tries to justify a cancellation of the social commitment after having cancelled it. Offering an excuse is a way of asking for being released from the obligation to perform what the agent is committed to. In fact, it can be argued that there are two kinds of excuses, depending on the type of responsibility that intends to be affected by the excuse: prospective or retrospective. An excuse is offered before cancelling the commitment in order to be released from the prospective responsibility acquired when committing oneself. Following our example above, Bob gives Ann a reason (that he has an important job interview the day after) with the goal of being released from his commitment (to lend his car to Ann). If Bob takes his car the day after in order to get to the interview, thus cancelling his commitment, he can offer this fact as an excuse as well. However, Ann cannot release Bob from his commitment, but can make him not accountable for violating her request[76]. In this case, excuses do not change the fact that Bob has unfulfilled a previous obligation, but serve to acknowledge that he has done so for good reasons.

There is a second way to justify the cancellation of a commitment, by arguing that the commitment was not valid, and thus the agent has not violated an obligation. This option exempts, rather than excuses, the debtor. Conditional commitments are an

74 Amongst others, see Kolodny and Wallace (2003), Owens (2006), and Gilbert (2011).

75 Watson (2005: 16); see also Watson (2009).

76 See §5.3.3 for a detailed explanation of how excuses affect responsibility attributions.

example of obligations that can be cancelled by certain external conditions. For example, Bob can accept Ann's request to lend his car to her, unless he is called for an appointment for a job interview; Bob is expecting this phone call, although he does not know for sure whether it is going to be the day after. If he does receive this phone call, he can cancel his commitment unilaterally, without violating any obligation he had towards Ann (although he ought to inform her about the cancellation). The external circumstances (the phone call) exempt Bob from his obligation. I will turn back to the relation between excuses, exemptions and responsibility attributions in §5.3.

CHAPTER 5: SOCIAL COMMITMENTS AND RESPONSIBILITY

The aim of this Chapter is to explore the relation between normative and empirical expectations, and attributions of responsibility. As I argued in Chapters 3 and 4, social commitments give rise to both types of expectations. On the one hand, social commitments are fulfilled on a regular basis, and this regularity creates empirical expectations about what agents will do. On the other hand, social commitments create social obligations, which are normative (or justificatory) reasons for actions that are socially acquired and acknowledged. By becoming socially committed, agents create normative expectations, and acknowledge being the target of such expectations. These expectations refer to what the agents should do.

Responsibility is a very broad concept. Moral and legal philosophy have provided many different definitions of what responsibility is, as well as different set of criteria an agent has to meet in order to be held responsible. The claim I will defend here is that responsibility can be understood in a non-moral and non-legal sense, as a relation between an agent and an outcome to another agent, who attributes responsibility. This relation is basically explanatory: an agent is responsible for an outcome if it is possible to explain the outcome in terms of the agent's authorship.

The structure of this Chapter is as follows. Section 5.1 analyses different concepts of responsibility, and argues for a normative notion of responsibility as attributability which precedes moral or legal responsibility. In §5.2, I examine two traditional criteria for attributing responsibility. Intuitively, an agent can only be responsible for what she has caused; I will present an overview of some problems related to causal responsibility. Then, I move to analysing the criteria regarding the agent's capabilities, both external (freedom

conditions) and internal (self-control and reason-responsiveness). In §5.3, I will present a contrastive account of causal explanations. The practice of asking and giving explanations is contextual to the background assumptions against which the explanation is required. I will argue that empirical and normative expectations play a fundamental role in explaining the causal relevance given to agents. I finally examine the relation between explanation and justification of action through the distinction between exemptions and excuses.

5.1. Kinds of responsibility: six different concepts

In a recent contribution, Vincent has explored the different uses that are given to the concept of responsibility, and how they relate each other[1]. Based on an example originally provided by Hart[2],Vincent explains the different uses of "responsibility" through the following story:

> (1) Smith had always been an exceedingly responsible person, (2) and as captain of the ship he was responsible for the safety of his passengers and crew. But on his last voyage he drank himself into a stupor, (3) and he was responsible for the loss of his ship and many lives. (4) Smith's defense attorney argued that the alcohol and his transient depression were responsible for his misconduct, (5) but the prosecution's medical experts confirmed that he was fully responsible when he started drinking since he was not suffering from depression at that time. (6) Smith should take responsibility for his victims' families' losses, but his employer will probably be held responsible for them as Smith is insolvent and uninsured. [3] (Vincent, 2011, 16)

While (1) refers to a virtue (presumably the virtue of living up to one's commitments), (2) is the kind of responsibility which Hart called role-responsibility, this is, the duties that are intrinsic to a specific social role; (3) would refer to outcome-responsibility, in the sense that Smith's actions led to an outcome for which she is responsible. The concept of responsibility as used in (4) denotes causation, and can be attributed to non-agents. In (5), it refers to the capability of the agent, this is, whether she has acted exercising full agency, or some internal or external constraints to her agency capability were at play. Finally, to

1 See also van de Poel (2011) for a similar taxonomy.
2 Hart and Gardner (2008: 211).
3 Vincent (2011: 16).

take responsibility as used in (6) is related to liability, and refers to the obligations the agent acquires in order to compensate a harm.

These multiple uses of the term "responsibility" can be divided into two general kinds. First, cases (1), (2) and (6) consist in the relation between an agent and future events; in this sense, they can be classified as different types of prospective responsibility[4]. On the other hand, cases (3), (4) and (5) refer to the agent's responsibility for a past event: they are types of retrospective responsibility[5].

On the other hand, (1), (2) and (6) can be analysed from the perspective of both legal and moral responsibility, whilst (3), (4) and (5) refer to a more general kind of responsibility. I will now explain these double classification.

5.1.1. Prospective and retrospective responsibility

Responsibility, similarly to agency, can be applied to past actions, or to future ones. As reflective agents, we are able to remember what we have done, and the reasons we had to do it. As planning agents, we foresee the consequences of our possible actions and choose what action to perform, according to our goals. The distinction between prospective (also called ex-ante and forward-looking) and retrospective (or ex-post) responsibility is often neglected in the literature[6]. Retrospective responsibility refers to the authorship of a present state of the world: it answers the question "who has done it?". When a detective interprets the evidence to find out who is Jones' murderer, she is looking for an agent who is retrospectively responsible for the death of the victim. A detective is looking for what caused the death of the victim, or more specifically, whose actions led to that death.

4 Virtue-responsibility (1) is both retrospective and prospective, for it means the virtue of living up to one's adopted and attributed role-responsibilities and obligations. In order to be considered a responsible person, in this sense, it is necessary to have previously fulfilled those commitments, and it is expected, because of the manifestation of that virtue, that the person will fulfil future ones. However, I will leave aside this concept of responsibility in the present Chapter.

5 Capability-responsibility can be applied to present or future states of affairs: it has to do with the agent's capabilities to act exercising full agency. However, in most cases, the agent's capabilities are evaluated after the outcome is produced, in order to know whether the agent is exempt because of lack of capabilities, so I will focus on this retrospective meaning.

6 See Richardson (1999); Anderson (2008); van de Poel (2011).

However, if the detective found that the cause of Jones' death was Smith's shot, this find only reveals that there is a causal relation between Smith shooting a gun, the bullet hitting Jones, and Jones dying, which would correspond to an attribution of responsibility as in (4) above. But it is not clear whether Smith's conditions are the right ones in order to hold her responsible for Jones' death in the sense of (3). Causal efficacy, I will argue, plays a controversial role in attributing outcome-responsibility. This is so because responsibility in the sense of (3) is an agential relation, and not a causal relation. The agent's relation to the outcome must be relevant to explain the existence of the outcome, but having causal efficacy and being a part of a causal explanation are not identical.

Prospective responsibility, on the other hand, is future-directed: it consists in being put (by ourselves or by others) in charge of a given situation. It has a teleological structure [7]. It is not about who actually does something, but about who should do it, or what should an agent do, as in cases (2) and (6) above. Taking responsibility means to become socially committed to perform the necessary actions in order to achieve or to preserve a state of affairs. This includes intending, promising, signing a contract, and acting accordingly to social norms, among others. The common element between these cases is that, by taking responsibility over an outcome, reasons for acting in ways that promote that outcome are created, as I have argued in the previous Chapter. There is a difference, however, between these different ways of acquiring a prospective responsibility over an outcome. The first three cases involve an agreement, and therefore the agent's awareness that she is being put under certain obligation. In some cases, the agent is aware of the responsibilities derived from a role or from a social situation. In other cases, the agent is unaware and, nonetheless, obligations still hold. Similarly, in some cases, these responsibilities are acquired and agreed explicitly, and in others, the agreement is tacit or even non-existent. Social obligations have to be socially acknowledged, but not universally acknowledged, in order to exist. This has a special significance when attributing retrospective responsibility, because the normative reasons for action the agent was

7 Birnbacher (2001).

supposed to have do not necessarily coincide with the motivational reasons for action the agent actually had for acting.

5.1.2. Attributability and accountability

Following Watson[8], when we attribute responsibility, we impute the outcome or the action to the agent, as her action, something she did, not merely at the causal level, but also at the level of authorship. An action, in this sense, discloses something about the agent's evaluative commitments: it reflects the agent's values. An agent is supposed to act for some reasons; and the weight she gives to the reasons she has is based on her values and previous commitments. As Bok argues [9], we evaluate those values compared with a standard (moral, legal -such as the legal fiction of a "reasonable person"- social roles, and so on) in order to examine her exercise of agency and to evaluate to which extent the outcome or the action shows the agent's reasons for action. The standard to which we compare someone's conduct consists of the normative and empirical expectations about her behaviour and reasons for action. Attributing responsibility, as I will argue in §5.3, affects the relevance of the agent's role in in the causal explanation of the outcome. On the other hand, making someone accountable is a social practice that relies on moral or legal norms, and violations of such norms can lead to being the target of reactive attitudes and associated practices. This distinction is exemplified by Ross metaphor of responsibility as a trial:

> The connection of responsibility with a trial shows that to be responsible for something can mean basically two different things corresponding to the two steps in the trial: accusation and judgement. In the first place being the person who can, when the situation demands, be rightfully accused (required to answer, give account); secondly, being someone who also satisfies the conditions of guilt and can therefore be rightfully sentenced. [10]

To rightfully accuse someone, then, would consist in finding out whether there is an agential relation to the outcome, and whether the agent is, for some reason(s), exempt.

8 Watson (1996).
9 Bok (1998: 183).
10 Ross (1975: 17).

On the other hand, to rightfully judge someone involves an evaluation of the conditions that make it correct, fair or reasonable to punish, blame or praise the accused. Given certain circumstances, both internal and external (as it happens with exemptions), the agent might be justified and therefore responsible in the first sense, but not accountable. Finally, an agent can be accountable, but offering an excuse mitigates her accountability. I will come back to the difference between exemptions, justifications and excuses in §5.3.3.

In the literature, this distinction is often made, although the terminology used can be quite confusing. Kutz[11] identifies responsibility-as-attributability with "responsibility", and accountability keeps its denomination. Pettit[12] refers to responsibility-as-attributability as "accountability" and, contrary to Kutz, identifies responsibility-as-accountability with "responsibility". Duff[13] draws the distinction between answerability and liability, which broadly correspond to Watson's concepts of attributability and accountability.

From Watson's point of view, accountability is an inherently social notion because it depends on the standards of certain communities. Attributability, however, is a judgement about whether the action discloses the agent's values, reasons and deliberative choices, this is, whether the outcome properly displays the hallmark of her authorship. As I stated above, I believe that this is also a social concept: we are responsible for something to someone, even in the attributability sense[14]. There are three reasons that support this claim.

First, when we judge that an agent is responsible, for we appreciate that the outcome reflects her agency in the appropriate way, evaluating her capabilities as an agent, we are adopting the stance of a judge, not the stance of the accused[15]. We take the evidence available as reasons to believe that the agent has displayed full agency, for

11 Kutz (2000).
12 Pettit (2007b).
13 Duff (2009a).
14 Duff (2009b).
15 Smith (2007).

example, and that no exempting conditions were at play. An agent has experiential and direct access to the reasons she had for acting, and the relevancy of external conditions in her choice. This observation leads to the other two reasons for claiming that responsibility is intrinsically social.

The second reason is that, from the point of view of a judge, we compare the agent's reasons for action and evaluative commitments with a standard of reasoning agent. This is why it is possible to be responsible for an outcome if it results from the agent's actions, even in cases in which the outcome was not intended nor foreseen by the agent. When an agent acts negligently, she has not acted as a reasonable person, this is, with reasonable care [16]. The legal fiction serves of reasonable person as a standard to compare not only someone's behaviour, but someone's deliberation mechanisms, evaluative commitments, and values.

Third, when we are evaluating the agent's relation with the outcome in order to find out whether she can be correctly "accused" of the outcome, the agent's reasons for action play an important role in the explanation of what is the relation between the agent and outcome. Not knowing how to swim exempts an agent from any responsibility for letting someone drowning, for example. So, the testimony of the accused is indeed relevant for our evaluation about the authorship of the action. Attributing responsibility to an agent makes it appropriate that the accused explains her actions, intentions, or reasons for action[17]. But acting does not commit us universally: not everyone is entitled to require such explanation. Thus, to adopt the stance of a judge requires certain level of authority over the agent's actions[18]. Sometimes, we lack of this authority: as it is

16 Keating (1996).

17 Oshana (1997).

18 In the case of moral responsibility, it can be argued that every moral agent has the authority to demand an explanation of the outcome to the responsible agent (in the case of retrospective responsibility) for moral wrongdoing, as well as to claim to be the holder of certain moral rights which put every other agent under the prospective responsibility of not violating these rights. In this sense, moral responsibility is not responsibility for something to a specific agent, but to every moral agent; this widens the scope of agents we are responsible to, but responsibility is still a social concept (although universally social). However, different moral communities have different levels of authority to ask for an explanation: for example, in Catholic communities, a priest would have the authority to demand an explanation for some morally blameworthy actions, while I would have not this authority.

commonly said, there are aspects of our life that are nobody else's business. Although we might be exercising full agency and intentionality, we do not owe an explanation to anyone but ourselves.

Attributability is conceptually (or logically) prior to accountability, although the practise of attributing responsibility usually takes place after something has gone wrong (or exceptionally right). Empirically, attributions of moral responsibility start with an outcome that we think morally significant, and then look for the author, and praise or blame her accordingly. Thus, I do not deny that reactive attitudes play a crucial role in attributions of responsibility: they usually trigger the social practises of blaming and praising. Contrary to the view that attributability precedes accountability, the Strawsonian tradition defends that responsibility actually works the other way around. In his well-known paper "Freedom and Resentment"[19], Strawson defends that moral responsibility can be understood in terms of the reactive attitudes (this is, feelings, emotions, and their associated practises) that we experience when we face moral wrongdoing. Strawson aims to criticise the claim that there is an external, rational justification for judgements of responsibility. As Watson puts it, in Strawson's view "there is not such independent notion of responsibility that explains the propriety of the reactive attitudes. The explanatory priority is the other way around: It is not that we hold people responsible because they are responsible; rather, the idea (our idea) that we are responsible is to be understood by the practice, which itself is not a matter of holding some propositions to be true, but expressing our concerns and demands about our treatment of another"[20].

The Strawsoninan account has had great impact on the debates about moral responsibility. Wallace[21] provided one of the most significant discussion and continuation of this tradition. Wallace makes a difference between someone being responsible and holding someone responsible, and claims that "conditions of responsibility are to be construed as conditions that make it fair to adopt the stance of holding people

19 Strawson (1962).
20 Watson (2004a: 222).
21 Wallace (1994).

responsible"[22]. Someone is responsible as long as it is fair to hold her responsible. Thus, Wallace introduces a normative aspect to the practices of having reactive attitudes and acting accordingly to them: it is possible to blame an agent in an unfairly, and therefore that agent would not be responsible (despite being held responsible). The fairness of the practice of holding responsible, following Wallace, is entrenched in social practises, so from his point of view, there can be no rational justification of judgements of responsibility.

This approach offers interesting observations about the practise of holding people responsible, but I believe that reducing the philosophical analysis of responsibility to the associated social practises misses the possibility of connecting theories of responsibility with theories of agency, which I believe to be interdependent. My aim in this Chapter is to argue for a concept of (non-moral) responsibility that, although dependent of social practises and standards, provides a justification for attributing responsibility that relies on agency and its relation with explanation and justification of action.

Furthermore, I believe that the Strawsonian account misses one important point. I have suggested, following Ross and Watson (amongst others), that responsibility is a two-staged process. However, I have focused on the difference between considering that an agent is responsible for an outcome, and judging that what she did was morally good or bad, or constitutes a criminal conduct, for example. But I think it is possible to draw a further distinction, which exceeds the scope of this work, although it is worth to be mentioned. I believe that making an agent morally or legally accountable (or liable) for something is different from judging that she deserves punishment or blame. As Smith[23] argues, the conditions for judging that someone is responsible (in Smith's paper, being a synonym of “blameworthy”), and the conditions for actively blaming her (or showing other reactive attitudes) are not the same. In the same thread of thought, Pettit points out:

> Holding someone responsible is distinct from just thinking the person responsible. Holding a person responsible requires thinking that the person is responsible, but it also involves a

22 Ibid., 15.

23 Smith (2007).

> further component. We think someone responsible when we think that the person satisfies conditions sufficient for being a candidate for blame or approval; we hold them responsible when we go one further step and actually blame or approve. [24]

Blaming can be inappropriate for several reasons that need not to undermine the agent's being responsible. I will develop further this distinction in §5.3.3, devoted to excuses and exemptions.

In conclusion, the concept of responsibility I will analyse in this Chapter refer to the proper agential relation between an agent and an outcome. It is based in Watson's concept of attributability, although it has a normative dimension, which is the standard against which the agent's conduct is evaluated and explained. In the next Section, I will explore two common intuitions regarding the criteria for attributing responsibility. The first intuition is that the agent need to fulfil certain criteria for having "fitness to be held responsible", in Pettit's terms. I will divide those criteria in three groups: freedom, control, and reason-responsiveness. Second, I will explore some problems of the (intuitively true) claim that responsibility requires a causal relation between the agent and the outcome. Finally, I will set forth which of those criteria have special relevance for causal explanations of the outcome in terms of agency.

5.2. Criteria for attributing responsibility

Defining responsibility is different from setting up the criteria that need to be fulfilled in order to be responsible for an outcome. Generally, two kinds of criteria are analysed. First, it is argued there must be some kind of causal relation between the agent and the outcome. This condition, however, faces controversial cases, such as omissions or overdetermination. Furthermore, there are competing accounts of what exactly a causal relation consists of. In addition, it seems that causal responsibility is not sufficient for grasping the complexity of responsibility of agents as a different kind of responsibility of non-agents, which would correspond to the mere causal effectiveness. This is why it is

24 Pettit (2007b: 173).

usually argued that, even if causality was a necessary criterion (which most of them seem to accept), the agent herself has to fulfil certain conditions. For instance, it has been argued that the agent has to have freedom to choose alternative options, to have a degree of control over her actions, or to be responsive to normative reasons.

Following Braham and van Hees[25], causality, freedom and agency conditions can be summarized as follows:

- Causal Relevancy Condition: There should be a causal relation between the action of the agent and the resultant state of affairs.
- Avoidance Opportunity Condition: The agent should have had a reasonable opportunity to have done otherwise.
- Agency Condition: The person is an autonomous, intentional, and planning agent who is capable of distinguishing right and wrong and good and bad.

Meeting the two first requirements, as I will show in this Section, can be problematic under certain circumstances. Philosophers have tried to adjust and refine the conditions under which causal and freedom requirements are met; specifically, the debates focus on cases in which, intuitively, the agent is responsible, although her causal efficacy is controversial; or, unbeknownst to her, she lacks from freedom of choice.

5.2.1. The agent's capabilities

Aristotle, in his Nicomachean Ethics, stated that responsibility requires both epistemic and "freedom-relevant" conditions. In order to be responsible, the agent has to bring about the action or the outcome voluntarily, understanding the relevant the particulars of that act, and freely causing the outcome[26]. Frankfurt[27] claims that these "freedom-relevant" conditions can be summed up in a principle, which he called PAP (for *Principle of Alternative Possibilities*): "a person is morally responsible for what he has done only if he could have done otherwise"[28]. Freedom conditions are external to the agent; when the lack of freedom is originated internally, it is denominated lack of self-control.

25 Braham and van Hees (2010: 7–8).
26 Raffoul (2010: chap. 1).
27 Frankfurt (1969).
28 Ibid., 829.

An agent needs to have at least some degree of self-control in order to be considered that she is exercising agency: compulsion undermines the agent's capabilities. The epistemic conditions envisaged by Aristotle have to do with the agent's capacity to deliberate and respond to (normative) reasons.

Control and freedom: Frankfurt-style cases

Frankfurt argues that the Principle of Alternate Possibilities is false, and therefore the controversy between determinism and free-will does not deal properly with the problem of the conditions for attributions of responsibility. Frankfurt's argument consists in showing that an agent can be responsible for her actions even in cases in which she has no possibility to act otherwise. Usually, the examples of agents lacking this freedom to act otherwise focus on cases in which the agent's capabilities are somehow undermined, such as compulsion and hypnosis. However, Frankfurt argues, there are other cases in which the lack of freedom cannot be understood as a lack of agency. He develops two thought experiments, of which the second of them has been largely discussed[29]. The first example consists in analysing the force of a threat. An agent is threatened to do something that she had already decided to do. Thus, the threat does not have any effect, for it does not influence the agent's choices. However, it is possible that, had the agent decided to do otherwise, the threat would have a deep impact on her, and forces her to act as ordered. In this case, it seems that the agent who decides to φ, is responsible for φ-ing, but not responsible if she acts against her will, and because of the threat. The difference lies in that, in the first case, the threat does not bring about the action; the agent does; while in the second case, it is the threat the cause of the action.

The second example has had a great impact in the literature about moral responsibility:

29 See, for a recent collection of essays on this topic, Widerker and McKenna (2006).

> Suppose someone –Black, let us say– wants Jones, to perform a certain action. Black is prepared to go to considerable lengths to get his way, but he prefers to avoid showing his hand unnecessarily. So he waits until Jones, is about to make up his mind what to do, and he does nothing unless it is clear to him (Black is an excellent judge of such things) that Jones, is going to decide to do something other than what he wants him to do. If it does become clear that Jones, is going to decide to do something else, Black takes effective steps to ensure that Jones, decides to do, and that he does do, what he wants him to do.8 Whatever Jones,'s initial preferences and inclinations, then, Black will have his way. [...]Now suppose that Black never has to show his hand because Jones, for reasons of his own, decides to perform and does perform the very action Black wants him to perform [30]

The means Black can take to prevent Jones from not doing whatever action Black is interested in can vary. Frankfurt considers a potion, hypnosis, and manipulating Jones' nervous system in order to simulate that Jones actually decides to do what Black desires. This last possibility has been the one chosen in the literature for developing what have been labelled "Frankfurt-style cases", although, as Vihvelin[31] argues, a kick or a punch would work equally well in the example. The interesting point in Frankfurt's examples is that, had Jones decided to φ (being φ "killing Smith"), Black does not show up in scene. So Jones would bear full responsibility for his φ-ing, although (unbeknownst to him) he had no real choice to avoid the outcome. Frankfurt's claim is that we must distinguish between being unable to do otherwise, and acting like this precisely because we are unable to do otherwise.

Critics to the validity of Frankfurt's argument can be divided in two kinds. First, it can be argued that Black's intervention, in whatever form it takes, affect some aspect of Jones' free will: the opportunity to act successfully according to our intentions and reasons. But Black cannot affect Jones' ability to deliberate, to choose according to his reasons and values, "the ability to be the agent-cause of one's actions" (Vihvelin 2008, p. 371). And it is this ability the relevant factor for attributing agency and responsibility. Second, it is controversial whether Jones' φ-ing under Black's manipulation constitutes a real case of "acting", or at least whether Jones killing Smith by his own will is the same particular event than Jones killing Smith under Black's intervention[32]. If the action in

30 Frankfurt (1969: 835–6).
31 Vihvelin (2008).
32 Van Inwagen (1978).

question consists in killing Smith, for example, it can be argued that, while without Black's intervention Jones would be actually killing Smith, the same is not true with Black's intervention: while Jones' movements may have caused Smith's death, it is not true that Jones has killed Smith[33]. Another similar point, defended by Álvarez[34], is that Black's capacity to make Jones to act is controversial: although he may be able to cause Jones to make certain body movements, it does not follow that Jones is acting.

Frankfurt's argument is interesting because it shifts the focus of the debate about responsibility from causal towards agential conditions that have to be met in order to attribute responsibility. This is, responsibility has to be founded in the role of the agent in bringing about the outcome, rather than the impossibility to do otherwise. The lack of alternatives is understood as an excusing factor, specially when the agent performed an action she did not want to perform. But it does not necessarily undermine agency. There are two main features that an agent has to possess in order to exercise her agency: displaying control mechanisms over her actions, and being responsive to normative reasons for action. These two conditions, which are closely related, ground attributions of responsibility.

Self-control and reason-responsiveness

Mental illness, compulsion and addiction are classic examples of loss of self-control. Under any of these circumstances, the agent's capability to control her own behaviour is undermined, in the sense that her deliberation mechanisms do not work as they should. In general, any form of loss of self-control affects the agent's capabilities *qua* agent. It is a widespread intuition that, in order to be attributable, an agent has to meet some minimal agency and rationality conditions. In Chapter 1, I analysed the relation between commitment, self-control and the stability of intentions; Chapter 2 and 4 were devoted to explore the rationality requirements of commitments. Thus, I will not expand those views

33 Larvor (2010).

34 Alvarez (2009).

again here; instead, I will focus on the relation between agential and rational requirements for fitness to be held responsible.

It is important to note that, in general, responsibility is not undermined when the agent does not fulfil a rational requirement (such as intending what she believes she ought to do, in the light of the reasons she has) or when she is weak-willed and drops a previous intention for no good reason. The standard view claims that responsibility is undermined when the agent is *not able* to respond to reasons, or to keep her intentions. Hence, those conditions refer to the agent's capabilities, rather than the agent's rationality failures. Children, for instance, are not attributable, because they do not meet those conditions. But an akratic agent is not exempt.

Following Fischer's influential account (both reflected in his solo works and in his works with M. Ravizza), "[a]n agent is morally responsible for performing an action insofar as the mechanism that actually issues in the action is reasons-responsive. When an unresponsive mechanism actually operates, it is true that the agent is not free to do otherwise; but an agent who is unable to do otherwise may act from a responsive mechanism and can thus be held morally responsible for what he does"[35]. The condition for responsibility attribution, rather than the possibility to act otherwise, consists in being responsive to reasons. For example, a person under hypnosis will act as commanded no matter what reasons she might have for or against acting in such way; thus, she is not exercising her agency, or put otherwise, her agential capabilities do not bring about the action, and she is exempt. The Principle of Alternative Possibilities requires that the agent has "regulative control" over her actions, which is a kind of control that allows the agent to take different paths of action. By contrast, Fischer and Ravizza[36] claim, regulative control is not necessary for attributing responsibility; instead, the agent needs to possess "guidance control" of her actions. Guidance control consists in acting in a way that our actions issue from our own reason-responsive mechanisms [37]. Guidance control has two

35 Fischer (2006: 66).
36 Fischer and Ravizza (2000).
37 Fischer (2010).

components: reasons-recognition (the ability to recognize the reasons) and reasons-reactivity (to choose in accordance with reasons that are recognized as good and sufficient). In the same line of thought, Wallace claims that, in order to be responsible, an agent must have reflective self-control, which is "the general ability to grasp and apply moral reasons and to regulate their behavior by the light of such reasons"[38].

Reason-responsiveness, thus, has two dimensions[39]. On the one hand, it is a *cognitive* capacity – the capacity to grasp reasons for acting. In the case of moral responsibility, the agent is required to grasp moral reasons, understanding the moral value of the available options. On the other hand, reason-responsiveness is a *volitional* ability: it requires that the agent is able to choose and act according to those reasons. Pettit summarizes those two aspects of reason-responsiveness into the following two conditions:

- Value judgment.—The agent has the understanding and access to evidence required for being able to make judgments about the relative value of such options.
- Value sensitivity.—The person has the control necessary for being able to choose between options on the basis of judgments about their value[40].

On the whole, reason-responsiveness conditions reflect the normative requirements of practical reasoning, on the one hand, and the standard control mechanisms of intentional agency, on the other. Failing to meet any of those requirements would undermine agency, and thus responsibility as well. I believe that these requirements are important in the theories of moral responsibility because they are consistent with the implicit assumption that, in order to be held responsible for an action, an agent has to be the author of that action. It does not suffice to have a causal impact on the outcome, but it is necessary a mark of authorship in order to justify attributions of responsibility. Watson's view follows this thread: in order to be held responsible, the agent's actions must flow from her

38 Wallace (1994: 155).

39 Nelkin (2008).

40 Pettit (2007b: 175). Pettit adds a third condition for fitness to be held responsible, which he calls "value relevance". It entails that the agent has to face a value-relevant choice; this is, agents cannot be held responsible for trivial decisions. Pettit's argument involves moral responsibility exclusively, and so this condition is not required in more general accounts of responsibility. Although the "value judgement" condition would also be only applicable to moral contexts, I believe it is a good description of the cognitive aspect of reason-responsiveness: the agent must be able to judge the value of her available options, and this, I believe, entails the grasping of reasons, either moral or non-moral.

evaluational system. This requirement suggests that the agent has to have caused the outcome in a way that it reflects her evaluation mechanisms and values, her "character". If the agent is nothing more than a link in a causal chain, and any other agent, under the same circumstances, would have done the same, then there seems to be no ground for attributing responsibility to her. This is consistent with the "agent-causation" view[41]: that an agent has to be able to modify the causal chain in order to cause the outcome as an agent, and not merely as an event in the causal chain. The same intuition holds in some views of self-control and strength of the will[42]: the agent must bring her actions in line with her resolutions, commitments, or judgements about what she ought to do. Was this an automatic process, the agent would not need to actively participate in it. The folk psychological intuitions about agency reflect this same intuition.

In the third Section of this Chapter, I will argue that it would be better to switch the focus from agent-causation to causal explanation in terms of the agent's reasons, intentions and evaluative mechanisms. This is, the causal explanation of the outcome must be such that the agential capacities of the agent have at least some relevance.

5.2.2. Causal responsibility

It is a strong and widespread intuition that a causal relation between an agent and an outcome is necessary in order to attribute responsibility to that agent[43]. Causal efficacy, however, is too broad to constitute the sole basis for holding someone responsible, for there are many things that we cause, and we are not responsible for all of them:

> To be agent-responsible for an outcome, the agent must be causally responsible for the outcome and the outcome must be "suitably reflective" of the agent's autonomous agency. There is much debate about what exactly determines when an individual is agent-responsible for something, but it's clear that one can be causally responsible for harm without being agent-responsible for it. [44]

41 O'Connor (1996).

42 Holton (2009).

43 See Weiner (1995); Sartorio (2007); Moore (2009).

44 Vallentyne (2009: 87).

For example, if I accidentally trip over a stone and fall, and as a result I break my leg, my body movements have caused I broke my leg (and in this sense I would be causally responsible), but I don't consider myself agent-responsible: the consequence of my actions (in this example, walking over a path with stones) was not and could not be foreseen by me, neither was it intended. The mental states and control capacity of the agent are indeed relevant, and will be discussed in the next Section. Strictly regarding the causal link between agents and outcomes, the controversy focus on what constitutes a causal relation, and what kinds of causal relation are appropriate to hold an agent responsible.

Offering an analysis of the metaphysics of causation would exceed broadly the scope of this work. My aim is to show the problems that arise from considering causation a necessary ingredient of responsibility. In the second part of this Chapter, I will defend that we should aim at causal explanations rather than causal relations, precisely because of the challenge that the four following problems represent to the causation condition. However, it would be odd to start analysing the problems without, stating at least a few general intuitions about causation. There are three main ideas that underlie the concept of causal responsibility, which derive from intuitions about the metaphysics of causation. First, causes are difference-makers. As Lewis puts it, "we think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it"[45].

This conception of cause as a difference-maker is central to counterfactual theories of causation, which, in a very simplified way, can be regarded as a derivatives from the claim: c is a difference-maker if, had c not occurred, e would not have occurred either. Thus, causation entails dependence between cause and effect. This leads to the second idea on causal responsibility, which apparently contradicts the first one: that in most cases, it is not possible to state that c has unilaterally, necessarily and sufficiently caused e. We can imagine plenty of examples in which an agent has causal relevance, thus "helping" to bring about or generate the outcome, although her sole action would not have been

45 Lewis (1973: 160–161).

sufficient, even if it is possible to assert that is was necessary, which is not always the case[46]. The paradigm of this kind of examples is collective action; however, regarding individual action, it is often the case that many other causes, of natural kind amongst them, have contributed to the outcome. Responsibility comes in degrees, as well as causal relevance[47].

Third, and deeply related to the control and intentional mechanisms that will be analysed in the next Section, causation is transitive. The goal of attributing responsibility is precisely find who or what is responsible for an outcome, amongst all the causal factors into play. These three intuitions (dependence, degrees of causation, and causal chain) underlie the concept of causal responsibility. I will now analyse three problems that those intuitions pose to the task of setting the conditions for attributing outcome-responsibility: the boundaries of the causal chain and deterministic scenarios, overdetermination cases, and the causal efficacy of omissions.

Causal chains and determinism

Causal relations are transitive: If *c* is a cause of *d*, and *d* is a cause of *e*, then *c* is a cause of *e*[48]. Hence, it is possible to trace all the events that have had causal influence on the actual income. Pulling the trigger caused the bullet to be fired, and the fired bullet caused the victim's death; hence, pulling the trigger caused the victim's death. Furthermore, my decision to pull the trigger caused that I pulled the trigger. Imagine that the context was such that it was highly likely that I decided to pull the trigger; for instance, I was provoked. Many factors might play a causal role in my taking the decision of pulling the trigger, and thus it is possible to track back the causal chain to those factors. Greene and Cohen illustrate this situation through a nice—although creepy—example:

> Let us suppose, then, that a group of scientists has managed to create an individual —call him 'Mr Puppet'— who, by design, engages in some kind of criminal behaviour: say, a

46 See Hall (2004) for an analysis of this distinction.

47 see Braham and van Hees (2009).

48 Hall (2004).

> murder during a drug deal gone bad. The defence calls to the stand the project's lead scientist: 'Please tell us about your relationship to Mr Puppet...'
> "It is very simple, really. I designed him. I carefully selected every gene in his body and carefully scripted every significant event in his life so that he would become precisely what he is today. I selected his mother knowing that she would let him cry for hours and hours before picking him up. I carefully selected each of his relatives, teachers, friends, enemies, etc. and told them exactly what to say to him and how to treat him. Things generally went as planned, but not always. For example, the angry letters written to his dead father were not supposed to appear until he was fourteen, but by the end of his thirteenth year he had already written four of them. In retrospect I think this was because of a handful of substitutions I made to his eighth chromosome. At any rate, my plans for him succeeded, as they have for 95% of the people I've designed. I assure you that the accused deserves none of the credit."
> What to do with Mr Puppet? Insofar as we believe this testimony, we are inclined to think that Mr Puppet cannot be held fully responsible for his crimes, if he can be held responsible for them at all. He is, perhaps, a man to be feared, and we would not want to return him to the streets. But given the fact that forces beyond his control played a dominant role in causing him to commit these crimes, it is hard to think of him as anything more than a pawn. [49]

It is not the case that Mr Puppet did not have control mechanisms, or freedom of choice, or that his mental states do not have causal force over his actions. What is stated is that the context shapes our values, what we take to be (good, reasonable) reasons for action, our goals. This is, our choices are not uncased causes: it is possible to trace what caused, either by influence or by determination, our decisions[50]. The difference between Mr Puppet and anyone else is that the initial conditions that explain his "configuration" as an agent are known. Those conditions track back to a group of scientists, instead of being the apparently random collection of events by which we have been influenced. Similarly, it would be possible to analyse the causes of the scientists' decisions, and so forth. Even if we limit the possible causes to agents, for our aim is to attribute responsibility, we face an infinite causal chain, where different agents pass the buck. In order to ascribe responsibility, the buck has to stop at some agent, and from the point of view of causality, it has to stop precisely because the special causal relation between the agent and the

49 Greene and Cohen (2004: 1780).

50 Although some authors argue that, in order to be responsible, agents must stand in a special causal relation to their actions, called by O'Connor "agent-causation", which is different from normal event causation O'Connor (1996). In his view, agent-causation is not determined by event causation; the agent can determine her goals and freely choose to act according to them. However, I will assume here that, either by influence, direct event causation, or through other causal relations, an agent's values and beliefs can be traced to its source.

outcome. When looking at the causal chain, we are looking for the sign stating "The buck stops here", such as the one U.S.A. President Harry S. Truman put in the Oval Office at the White House, while governing.

Thus, when we look at the causal chain leading to the outcome we want to attribute responsibility for, the agent seems to be just a link in a much larger chain. If all the events are caused by certain initial conditions and their changes due to physical laws, then the agent may indeed be such link in the chain. Causal determinism states that every event is a link in the causal chain, so every cause is an effect as well. However, intuitively, one can only be responsible for an outcome, which is the effect of the agent's actions, only if the agent is responsible for those actions as well. Then, by analysing the causes of the agent's actions, we pass the buck backwards and relieve the agent from her responsibility. The buck would stop by showing that the agent acted freely (this is, she could have chosen not to act). Determinism and free will, enter into conflict[51].

Overdetermination

Overdetermination occurs when there are at least two distinct and sufficient causes for the same effect[52]. Take the following example, called "Two Assassins":

> Sam and Jack each want to assassinate the mayor of their town. Each has his own nefarious and no doubt misguided reasons for wanting the mayor dead. They act entirely independently of each other - they do not even know about the other's existence. Each of them deliberates and acts in a way which apparently makes him morally responsible for his actions - neither is compelled, coerced, deceived, crazy, manipulated electronically, hypnotized, and so forth. Both Sam and Jack go to a city council meeting at the town hall, and simultaneously pull the triggers of their guns. Their bullets strike the mayor at the same time, and the processes leading from each bullet's hitting him to a sequence of life-threatening biological events are similar. Each bullet's hitting the mayor in the way it does is sufficient for the mayor's being killed by it. Moreover, the situation is such that neither individual could prevent the other from shooting and killing the mayor - perhaps each is wearing a bullet-proof vest and other protective equipment. We could add further specifications which would make it the case that neither individual ever had the opportunity to prevent the other from being in a position to shoot and kill the mayor, though the details

51 See for an overview on this problem French, Wettstein, and Fischer (2005); Fischer (2007); Kane (2011).

52 Funkhouser (2002).

would only clutter the discussion here. [53]

Overdetermination examples involve some sort of unlikely coincidence at some point along the causal chain leading to the outcome, in which each of the two independent and different events would cause the outcome without the intervention of the other[54]. Typical cases are the one described in the example "Two Assassins", two rocks hitting simultaneously a window and breaking it, firing squads, and so on. These examples pose a challenge to the counterfactual theory of causation, because, had any of the assassins not shoot, the victim would have died anyway. Besides, both assassins seem responsible for the victim's death, at least in the sense of outcome-responsibility.

A broader version of overdetermination does not require that there are two distinct causes, but only that would one of the causes not have acted, the outcome would have been the same. In the the example of the two assassins, if Sam shot the victim a few seconds before Jack did, Jack's bullet would impact in a dead body, not causing his death. Therefore, Jack would not be causally responsible for the mayor's death; however, had Sam not shot, Jack would have caused his death.

Finally, a form of overdetermination occurs in large-scale outcomes. For example, although we are all individually responsible for global warming, my individual and particular contribution does not add a significant cause, and is indeed neither necessary nor sufficient for causing the outcome. Parfit argued against what he considered a mistake of moral mathematics (the fifth of them): "If some act has effects on other people that are imperceptible, this act cannot be morally wrong because it has these effects"[55]. In an example called "The Harmless Torturers", Parfit describes a group of torturers who increase each victim's pain in imperceptible ways, but the resulting increase of pain after all the torturers have done their part is indeed perceptible. Each contribution, as in the

53 Fischer (1998: 216).
54 Sartorio (2006).
55 Parfit (1984: 75).

case of global warming, has a cumulative impact, but cannot by itself cause the outcome[56]. This aspect of causal responsibility will be discussed in Chapter 7.

Omissions

Agents are often held responsible for something they have not done, such as preventing an outcome[57]. Omissions are problematic because the relation between the agent and the outcome is not evident: there has to be a link of some kind that connects agents with outcomes. However, it is not clear whether that link is necessarily causal, or on the contrary, causal effectiveness is not necessary for responsibility. Let's suppose that omissions have causal effectiveness, and thus some omissions indeed cause some state of affairs to be the case. From the point of view of counterfactual theories of causation, had you watered your plants, they would have not died, hence your not watering the plants is the cause of your plants' death[58].

The first problem for "absence causation" is that there is an infinite number of things we do not do; the concept of omission needs to be narrowed in order to make it more specific than "inaction" or "non-action"; otherwise we would be causally responsible for all the things we fail to do:

> There are various different conceptions of omissions. One way of classifying them distinguishes wider and narrower conceptions of omissions. On the wider conception (which may not link up closely with ordinary usage), whenever a person does not do something, X, he fails in the relevant sense to do it, and he omits to do it. Thus, we are all now failing to stop the Earth's rotation (and omitting to stop the Earth's rotation). Omission to do X (according to the wide conception) need not require explicit deliberation about X, and it need not require the ability to do X. [59]

Restricting the set of omissions can be achieved through various strategies. For example, inactions (which would be the widest set of non-events we do not cause) can be narrowed

56 Attfield (2009).

57 As Dowe (2001) claims, omissions can be found in three different scenarios: when they are act as cause (A's not φ-ing caused outcome X), when they are a consequence of one's action (A's φ-ing prevented X, in the sense that caused X not to occur), or they can be both (A's not φ-ing prevented outcome X). Here, I will focus on the first and last cases, in which the omission has causal efficacy.

58 Dowe (2009).

59 Fischer (1997: 46).

through requiring the ability to act (in Fisher's example, we are failing to do something that we actually are not able to do), or requiring that the omission was the result of a deliberative process: we choose not to act. However, those restrictions refer to the agent's capabilities and intentionality. It is difficult to narrow omissions by their causal effectiveness. And thus, accepting the causal effectiveness of omissions can lead to counterintuitive consequences, at least from the point of view of counterfactual theories of causation. Not only the scope of omissions has to be narrowed to avoid dealing with an infinite number of inactions, but also the number of agents who, counterfactually, could have avoided the outcome through acting. The Queen of England problem was first proposed by Sartorio, in the following terms:

> If we were to say that my failure to water a plant that I promised to water is a cause of its death, then we would probably also have to say that the Queen of England's failure to water the plant is a cause of its death (because it is also true of the Queen of England that, had she watered the plant, the plant would have survived). [60]

Hence, the problem is twofold: not only there is an infinite number of actions an agent has not performed, but also there is an infinite number of agents who have not performed an action that, counterfactually, would have avoided the actual outcome. These are two different puzzles, although they are closely related. They both refer to the appropriate relation between agents and outcomes, in which responsibility can be founded. The theories that deal with the causal relation between actions and omissions approach this problem from a variety of perspectives. For instance, it is argued that omissions would be indeed actions performed by an agent[61]; some of the strategies involve to provide alternative accounts of causation in order to allow certain omissions to cause an outcome[62]; to include normative considerations in our understanding of causality[63]; or to delimit the set of agents who are causally related to the outcome by discriminating those

60 Sartorio (2004: 322).

61 See Boniolo and De Anna (2006); Sneddon (2006); Clarke (2010).

62 See Menzies (2004); Longworth (2006); Sartorio (2004); (2007); (2009).

63 McGrath (2005).

who had not a serious possibility to act[64], which would explain why we do not hold the Queen of England as causally responsible for the plants' death. Finally, it is possible to accept that Sartorio (in her example) and the Queen of England stand in the same causal relation with the death of the plant – either a genuine causal relation[65], or none at all[66]. I will take up again the Queen of England problem in §5.3.2.

5.3. Responsibility, expectations and explanation

The aim of this Section is to analyse the link between responsibility attributions and causal explanations. My claim is that attributing responsibility to an agent for an outcome requires that the agent plays a role in the explanation of the outcome. This Section is divided as follows. First, I will argue that causal explanations are context-sensitive, and thereby they have (at least implicitly), a contrastive form. Contrasts are given by the relevance attributed to different facts, which is conferred by their salience with regard to the background assumptions of the agent who is providing the explanation. I will clarify the relation between explanation and responsibility through the analysis of some examples found in the literature of responsibility attributions. Finally, I will explore the relation between explanation and justification of action by showing how excuses and exemptions affect attributions of responsibility. The distinction between excuses and exemptions, I will argue, supports the claim that responsibility is a two staged process (following Ross' metaphor, accusation and judgement), as I claimed in the first Section of this Chapter.

5.3.1. Responsibility and causal explanations

When we hold someone responsible for a plant's death, we are asserting that the plant died *because of* that agent actions or omissions. Similarly, when we make a doctor

64 Woodward (2003).
65 Schaffer (2010).
66 Moore (2009).

responsible for a patient's death, we state that the patient died *because of* the doctor's actions: by negligence, lack of expertise, or by poisoning the patient with hemlock, or stabbing him. Regardless of the means (although not independently of them), there is a causal link which allows for explaining the patient's death through the doctor's actions, not only through his behaviour, but also through his agential conditions, such as his capabilities or level of expertise. This is, we link the agent and the outcome in a causal explanation. As I have showed in the previous Section, this can be highly problematic depending on what is accepted as a true causal relation –infinite causal chains, omissions and overdetermination seem to challenge the assumption of the necessity of causing an outcome in order to be responsible for it.

Context-sensitivity and contrastive causation

Hence, it is necessary to bridge the gap between causality, understood as a physical connection or as a transmission of some property[67] and causal explanations. The distinction between causal relations and causal explanations, stated by Davidson, can be understood in a variety of ways. Causal explanations are context-sensitive, in the sense that some facts are more explanatorily relevant than others amongst all the facts that are causally related to the outcome to be explained. The context-sensitivity of causal explanations may have different justifications[68]. First, it can be argued that the concept of "cause" is ambiguous and it sometimes refers to causal relations, and in other cases it means "causally explains" [69]. The difference between both is a case of context of use; and the correctness of causal explanations depends on the relevance or importance of some causes amongst others in that context.

Second, the concept of cause may be unambiguous but, given that explanations are given in a conversational context, this same context provides the cues for highlighting

67 See Dowe (2000); (2004); Salmon (1998).
68 Schaffer (forthcoming) (forthcoming).
69 Davidson (1967).

some causes amongst all of them. This point of view locates the origin of context-sensitivity in conversational pragmatics. For example, an event –an aeroplane crash- is the result of several converging conditions, which are its causes. However, different causal explanations can be given. For an engineer, the mechanic causes are more relevant, while a meteorologist highlights how the atmospheric conditions affected the engine.

Third, contextualism claims that the context sensitivity of causal explanations is in part due to semantic constrictions. Following Hart and Honoré [70], "the contrast of cause with mere conditions is an inseparable feature of all causal thinking, and constitutes as much the meaning of causal expressions as the implicit reference to generalization does" (1985, p. 12).

For all the three approaches on context sensitivity, it is widely accepted that causal explanations highlight some causes as more informative, relevant or important than some other causes, even if those other causes are necessary conditions for bringing about the outcome:

> What explains an event, on the causal approach, is something about the causal process that produces the event. It seems that not every causal fact is explanatory, however […]. [A] complete causal account of explanation must include a criterion for assessing explanatory relevance. [71]

The correctness of an explanation depends on the explanatory background against which we demand or offer the explanation[72]. The background makes it appropriate to stress certain factors as specially relevant for explaining the outcome. This is so because an explanation is the answer to a question – usually, a question about why something is the case.

---

70 Hart and Honoré (1985).

71 Strevens (2004).

72 My argument in this Section is that background assumptions, specially the agents' normative and empirical expectations about what will / should happen, conform the context against which the explanation is required and given, and provide reasons to consider certain factor as more relevant than others. However, this claim is compatible both with the view that defends that only pragmatics can explain this context-sensitivity, and also with the view that at least some semantic properties are needed to explain relevance. I will not try to clarify here this debate, or take part into it, for it exceeds the scope and aim of this Thesis. Thus, I will assume that any of those two points of view might be correct.

Attributing responsibility, I will argue, consists in judging that the outcome can be explained in terms of the agent, either through her intentional actions, her evaluation mechanisms (including her reasons for action) or the deliberation processes underlying her choices. While an unintentional omission can be hardly described as an action, for example, the agent's lack of responsiveness to certain reasons she is supposed to have (as, for instance, in cases of negligence, in which the agent is fully capable, but her reasoning processes differ from the normative standard of a reasonable person) can be the basis for situating her in the causal explanation of the outcome. The criteria to highlight an agent as a relevant cause of an outcome depend on the specific explanatory background against which the explanation is requested.

The traditional structure of a causal explanation is binary: it takes the form c causes e. However, this relation does not explicit why c_x – amongst c_1, c_2, c_3... c_n – is explanatorily relevant for e. Causal explanations can be contrastive in three different ways. First, the contrast can be situated in the effect side of the explanation[73]. Hence, causal explanations would take the form "c causes e rather than E'". The alternative effect (E') is the background against which c is relevant for explaining e. For example, the fact that Ben is hungry (c) causes that Ben is eating an apple (e) rather than not eating anything at all (E'), but the same fact does not explain why Ben is eating an apple (e) rather than a pear (E'') (in this case, the cause should point out something about Ben's preference for apples over pears, for example).

Second, the contrast can be placed on the side of the cause: "c, rather than C', causes e". An event has only causal relevance compared to other events. For example, when asking whether moderate smoking causes cancer, the response can be either affirmative or negative, depending on what alternative causes are under consideration:

> The solution to this puzzle is to deny that there is any such thing as the causal relevance of moderate smoking for lung cancer. [...] Relative to heavy smoking, it is a negative cause of (prevents) lung cancer; relative to abstaining, moderate smoking is a positive cause of (causes) lung cancer. [...] Relations of positive or negative causal relevance only hold relative to

73 See Van Fraassen (1980).

specific alternatives. [74]

A third approach, defended by Schaffer[75], consists in arguing that causation is contrastive on both sides, cause and effect. Therefore, causal explanations would take the form "c, rather than C', causes e, rather than E'". It is important to note that C' and E' do not need to represent not-c or not-e necessarily. For example, it is possible to explain Ben's picking an apple rather than not picking an apple by appealing to the fact that Ben is hungry rather than not being hungry; however, explanations can also be constructed alluding to other contrasts. Ben's decision to skip his strict diet rather than following the doctor's recommendations explains why he has eaten a piece of cake rather than eating an apple. C' and E' make the background against which the explanation is given explicit.

I believe this approach to causal explanations is the most suitable to responsibility attributions for two reasons. First, when making an agent responsible, it is claimed that this agent is more relevant to the explanation of the outcome than other factors. For example, in the context of a military chain of command, some agents (for instance, those with the lowest rank) can be excused under some circumstances, even of they have participated in the production of the outcome. Furthermore, making someone responsible entails to argue that this someone, rather than some forces out of her control (making the agent exempt; see §5.3.3 below), can causally explain the outcome. For example, Sarah's Obsessive Compulsive Disorder, rather than her belief that her hands are full of bacteria, causes her to wash her hands every so often, instead of resisting the urge. The second reason for preferring the contrast view of causation is that it allows for degrees of causation or causal relevance, which can be translated into degrees of responsibility[76].

Explanation, background assumptions and expectations

74 Hitchcock (1996: 402).

75 Schaffer (2005); (forthcoming).

76 For an non-contrastive account of causation that also gathers the problem of degrees of causation, see Braham and van Hees (2009).

A causal explanation consists, in general, in answering a question about why something is the case. A correct explanation does not require to mention all the causal relations and factors into play; some of these factors are selected and highlighted, depending on their relevance for the context of the explanation. The context, or background assumptions, include both the inquirer's and the explainer's beliefs, presuppositions, and expectations. They function as a *ceteris paribus* clause, stressing the relevance of certain factors when all the other background conditions remain constant[77].

As I argued above, responsibility attributions, as long as they entail a causal explanation of the outcome for which an agent is being held responsible, take the form of double contrasts, in which C' and E' are the background assumptions against which the explanation is required. Normative and empirical expectations play an important role as background assumptions that contrast with how things have actually developed.

In order to be held responsible, an agent has to have played a role in the causal explanation of that outcome. This requirement does not entail that the agent needs to have caused the outcome, in the narrow sense of causation. As I showed in the first Section of this Chapter, the requirement for the agent of having strictly caused the outcome is too strong, and makes many instances of responsibility attributions problematic. In this sense, responsibility is similar to authorship: to attribute authorship of an action or state of affairs consists in explaining that action or state of affairs appealing to the author, taking as evidence the marks of her authorship.

In a recent contribution, Björnsson and Persson propose that being morally responsible involves the idea of the outcome being explained by the agent's motivation. They develop the "explanation hypothesis", which takes the following form:

> Explanation Hypothesis: People take P to be morally responsible for E to the extent that they take E to be an outcome of a type O and take P to have a motivational structure S of type M such that GET, RR and ER hold:
> - General Explanatory Tendency (GET): Motivational structures of type M are significant parts of a reasonably common sort of explanation of outcomes of type O.

77 Schweder (1999).

> - Reactive Response-ability (RR): Motivational structures of type M tend to respond in the right way to agents being held responsible for realizing or not preventing outcomes of type O.
> - Explanatory Responsibility (ER): The case in question instantiates the right sort of general explanatory tendency: S is part of a significant explanation of E of the sort mentioned in GET. [78]

The GET condition constrains the relation between motivation and outcomes to what is normal or expected. RR refers exclusively to attributions of moral responsibility. The authors argue that the motivational structures that explain the outcome have to be modifiable or influenced by the reactive attitudes of praise or blame towards the agent; if those motivational structures cannot change, or are not affected by reactive attitudes, there is not any reason to blame or praise the agent. Lastly, ER is the particular stance of the more general condition GET.

Although Björnsson and Persson's analysis is concerned with moral responsibility, the two remaining conditions above (GET and ER) are necessary for attributions of responsibility, which would be, as I have argued in the previous Section, a condition for attributing moral or legal responsibility. I believe that the analysis provided by Björnsson and Persson is correct; I will now turn to the relation between background assumptions, expectations and explanations.

The claim I want to defend here is that, in order to hold an agent responsible, that agent has to play a role in the explanation of what she is being held responsible for. This is, she (her actions, intentions, reasons for action...) has to have some explanatory relevance in the production of the outcome. This relevance is given by the context against which the explanation is required. Background assumptions discriminate the important or relevant causes from other candidates, through contrasting its role in the production in the outcome, and through evaluating alternative outcomes. Double contrast models of explanation gather these features. Both normative and empirical expectations shape our background assumptions, altogether with our background beliefs about how the world is and behaves.

78 Björnsson and Persson (2012: 5).

Mechanistic explanations provide a useful analogy of the underlying process of discrimination and relevance attribution to causes and effects[79]. Following Barros[80], it is common to most accounts about mechanisms and mechanistic explanations that "a mechanism that explains a phenomenon includes things in the world (parts or entities) and things that these things in the world do (interactions or activities)" (2011, 8). When the mechanism works as expected, explaining the production of the outcome would require to mention what parts it has and how they interact. For example, to explain why a car moves when the driver activates such and such devices would include a description of the parts of the car involved, and how they relate to each other. However, mechanisms, in a broad sense, can be more vague and flexible. I have set up a (quite simple) mechanism to get sure I wake up on time in the mornings: I have two alarm clocks, one of them next to my bed, and the other one outside. The outside clock is set to the time I really need to wake up if I do not want to be too late, while the closest alarm is set half an hour before, because I can easily reach it and decide whether to stay 10 minutes more. I do not trust myself half asleep, and so the second alarm is my *commitment technology*, in case I turn the alarm off. It would be unnecessary to mention this mechanism to explain why I got up at 8; I would probably mention the reasons I had to wake up at 8 (and those same reasons justify to set up the alarm). If I am late to an appointment, and I get asleep, the specific failure of my mechanism would explain why I could not get up on time. If the closest alarm did not sound, the second one would awake me, but unfortunately too late for arriving on time. If the second alarm does not sound, I would probably believe that I have time to sleep 10 minutes more, and stay in that loop until noon. The alarm clock's failure to ring (an absence) explains why I got up late. Or, on the contrary, I could have set up any of the alarms too late. In this case I am responsible for not arriving on time to my appointment: my actions explain why the mechanism has failed.

79 Mechanisms also have some parts which have specific functions. They enable or produce that the mechanism works properly. This is also a nice analogy of prospective responsibility, understood as the duties and obligations an agent has.

80 Barros (2011).

I believe that many scenarios of responsibility attributions used in the literature as examples, or with experimental purposes, are analogous to mechanisms. For instance, in Frankfurt-style cases, as explained above, Black and Jones are parts of the mechanism that produces the outcome (Smith's death). In the experiment conducted by Knobe and Fraser[81], which I will analyse below, the group of professors, administrative staff and receptionists form a mechanism which has a causal impact on the shortage of pens.

My claim is not that all causal explanations are mechanistic, or that attributions of responsibility require a mechanistic explanation. My aim is to note the analogy between two forms of reasoning: on the one hand, when a mechanism fails, we check the correct functioning of each part, and a malfunction explains the general failure. On the other hand, when we attribute responsibility, we usually want to explain an unexpected outcome; if the outcome is expected, an explanation is rarely required. Of course, it is possible to explain an expected outcome and holding agents responsible for it. Imagine that Bob has broken his leg. He goes to the hospital, a doctor examines the leg, Bob has his leg x-rayed and has his leg put in plaster. After four weeks, Bob goes back to the hospital and has his plaster removed. The doctor then confirms that his leg is cured. If we aim to explain why Bob's leg cured, we would probably have to tell this story, or, depending on the context, a story about how bone regeneration works. It does not follow that the medical staff (and Bob through his own care) is not responsible for the recovery of the leg. They took the decisions and performed the actions which lead to the outcome. But this explanation would better answer how things happened, rather than why things happened. In this sense, when a mechanism works as expected, an explanation of its product usually consists in a description of the mechanism.

Moreover, the correct functioning of a mechanism is given by certain descriptive norms, which refer to the normal behaviour of an agent or object:

> Sometimes 'norm' means what people commonly do in certain situations, what constitutes 'normal' or 'regular' behavior. This notion of regular behavior differs in important respects both from a shared habit and from what people believe ought to be done, what is socially

81 Knobe and Fraser (2008).

> approved or disapproved. [82]

Following the example above, if my alarm clock does not ring and I wake up late, I may wonder why it did not ring: it should have rung. This "should" does not requires that the object (the alarm clock) has reasons to ring. It means that I have reasons to believe that, unless some abnormality happens, the alarm will ring. It is an empirical expectation; however, it is put under the form of a normative expectation. It is normal for X to φ if X is supposed to φ[83]. In the case of my alarm not ringing, something has gone wrong, because it did not do what it was supposed to do. Maybe I did not set the alarm properly, or the alarm clock has no batteries, it is broken, or my flatmate took it. The mechanism I use to wake up on time is not working well. And, in this sense, our expectations about what things will happen, depending on the normal, typical or regular behaviour of all their dependencies, generates a kind of expectation which relies on descriptive rules, which is empirical in a sense (what will happen) and normative in another (what is the correct or normal functioning). I think this clarification is important in attributions of responsibility because unexpected outcomes usually take the form "this shouldn't have happened" (if everything worked correctly, or everyone had behaved as expected), but the "should" here does not imply that a moral rule has been violated.

To sum up, attributions of responsibility can take place for any expected or unexpected outcome; however, from a pragmatic point of view, explanations are usually requested when the outcome is unexpected. In the case of moral responsibility, as pointed out above, blame and praise are logically subsequent to attributing outcome-responsibility, in the sense that an agent who is not outcome-responsibility cannot be blamed or praised for her actions, omissions, choices, and so forth, if there is not any connection between her and the outcome. Both blameworthiness and praiseworthiness of actions are attributed for unexpected outcomes[84]. An action that would deserve blame if unexpected (such as

82 Bicchieri (2006: 29).

83 McGrath (2005).

84 Wallace, as well, places expectations in the centre of his theory about responsibility. However, he argues that the relevance of expectations lies in their connection with the moral sentiments:

killing someone) can, if expected, no to be blameworthy. As I will argue below, cases of provocation can make the agent exempt. On the other hand, an action that would deserve praise if unexpected can, similarly, not to be praiseworthy if expected. For example, not slapping someone does not usually deserve praise; however, not slapping someone who has provoked us to unreasonable limits can be an action worth of praise.

5.3.2. Two examples

At the beginning of this Chapter, I drew the distinction between prospective and retrospective responsibility. Prospective responsibilities (such as social commitments) are a source of normative expectations, as I showed in Chapter 4. Attributing responsibility does not only consist in evaluating what an agent has done, but to evaluate it under the light of what she should have done, appealing to a normative standard, which is not necessarily moral. In this Section, I will analyse how expectations and background assumptions serve as contrast factors in double sided explanations, which are the basis for attributions of responsibility. To serve this purpose, I will examine two examples of responsibility attributions in the literature. The first of them is taken from an experiment by Knobe and Fraser, and involves normative and empirical expectations which do not derive from an explicit commitment of the agents. The second example is the Queen of England problem, presented in §5.2.1. It contains an explicit social commitment, which is the basis for the normative and empirical expectations forming the context against which an explanation is required.

> [E]pisodes of guilt, resentment, and indignation are caused by the belief that an expectation to which one holds a person has been breached; the connection with expectations gives the reactive emotions common propositional objects, tying them together as a class. Once this interpretation of the reactive emotions is in place, we can draw on it to account for the stance of holding people morally responsible. Wallace (1994: 12).

From this point of view, expectations play a role in the origin of the reactive attitudes towards an agent. I do not intend to argue against this claim: on the contrary, I believe emotions and expectations are deeply connected (see Chapter 3). However, I do not believe that neither moral or non-moral responsibility conceptually require reactive attitudes; thus, the role I attribute to expectations is very different from the role attributed in Wallace's theory.

Social norms: "Shortage of Pens"

The first example is taken from an experimental study by Knobe and Fraser[85]. They offer an explanation of the differences observed regarding responsibility attributions based on moral considerations. This idea has been given some attention in the literature. Alicke[86] conducted an experiment that showed that moral considerations of the reasons for violating a norm have an impact on the causal relevance attributed to the violator: "[w]ith causal necessity, sufficiency, and proximity held constant, the more culpable act was deemed by subjects to have exerted a larger causal influence"[87]. Joshua Knobe argues that "causal attributions are not purely descriptive judgements. Rather, people's willingness to say that a given behaviour caused a given outcome depends in part on whether they regard the behaviour as morally wrong"[88].

I believe that a further exploration of the background assumptions (not necessarily moral assumptions) can provide a better understanding of the causal attributions observed in the results. In the conducted experiment, the following narration is offered:

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own.
> The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist has repeatedly emailed them reminders that only administrative assistants are allowed to take the pens.
> On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later that day, the receptionist needs to take an important message... but she has a problem. There are no pens left on her desk. [89]

The subjects participating in the experiment were asked to say their degree of agreement with each of these propositions: "Professor Smith caused the problem" and "the administrative assistant caused the problem". Participants were prone to agree with the claim that Professor Smith caused the problem, while disagreed with the statement that

85 Knobe and Fraser (2008).
86 Alicke (1992).
87 Ibid., 370.
88 Knobe (2006: 62).
89 Knobe and Fraser (2008: 443).

the assistant caused the problem. However, Knobe and Fraser argue, both behaviours are equally frequent; the only difference stems from the different moral value of each action (it is wrong, we assume, to violate the norms). Therefore, they conclude that moral judgements affect the process by which we make causal claims.

Their conclusion has been revised by Roxborough and Cumby[90], who argue that the experiment lacks from a differentiation of a crucial factor to attribute a role in a causal explanation: the typicality (or atypicality) of events. The authors conducted another experiment, with the same scenario, but introducing a variation on how typical the rule-following and the rule-violation behaviour was. They conclude that the typicality or atypicality of a behaviour affect causal judgements. Furthermore, a variation in the typicality of the competing causes (in this example, the administrative staff behaviour) affects attributions of causal relevance to Professor Smith, even when the typicality of his behaviour remains constant.

The conclusions of this second experiment can be easily translated into a double contrast model of explanation, in which the typicality (this is, what is expected as normal) of the contextual elements is evaluated in order to attribute a higher relevance to one of the elements. Knobe and Fraser argue that the causal claim made by the participants in their experiment is thus the following: Professor Smith taking one of the two remaining pens (c), rather than the administrative assistant taking one of the two remaining pens (C'), causes the receptionist not being able to write down a note (e), rather than being able to write it (E'). Under their interpretation, this result can be explained through the moral value given to (c). Following Roxborough and Cumby, by varying the frequency of (c) and (C'), the relevance given to those two possible causes vary as well. I will now develop this second experimental results under the light of the attributer's expectations. But before that, I will make a small remark. I believe that the fact that Professor Smith has a proper name, while the administrative assistant and the receptionist do not, has the possibility to affect attributions of responsibility. For example, people tend to attribute

90 Roxborough and Cumby (2009).

more responsibility to agents in concrete scenarios than in abstract ones[91]. For this reason, and also for the sake of brevity, I will name the administrative assistant "Ann", and the receptionist will be named "Ruth". Professor Smith will be called only "Smith".

What has to be explained, then, is why Smith's taking a pen is more relevant than Ann's taking another pen, on the one hand, and what kind of assumptions we are making when we take it for granted that the lack of a pen is the key difference between being able to grab a message and not being able to do so. I will start with the second part.

Neither Smith's nor Ann's actions explain why Ruth was not able, for example, to write the message in her computer, or to memorize it. The explanation-seeking question, borrowing Schweder terminology, is what caused Ruth not being able to write down a message with one of the pens located in the normal place of pens. This question is assuming that, under normal circumstances, messages are written with those pens, and no other means are used to take messages. After all, if it were expected that Ruth types all her notes on her computer, the shortage of pens would not causally affect taking important messages. Thus, the inquirer's expectation is that the messages are taken by Ruth with those pens, and that messages should (in a descriptive sense) be taken with pens. These assumptions justify to highlight what caused the lack of pens as the cause of why Ruth was not able to write down a note.

On the cause side, the contrast is given by the scenario proposed in the experiment. In Knobe and Fraser's experiment, Smith's action, rather than Ann's, causally explains Ruth's problem. Smith is not normatively expected to take pens, while Ann is. However, if the typicality of Smith's and Ann's actions vary, so do their causal relevance, as Roxborough and Cumby point out. Thus, both empirical and normative expectations play a role in attributing salience to one of the options above the other. A violation of expectations of both kinds actually generates the search for an explanation. As Schweder suggests, "usually, what gives rise to an e[xplanation]s[eeking]-question is some event that

91 Nichols and Knobe (2007).

is considered surprising or in need of explanation by someone, the inquirer"[92]. Intuitively, were both professors and administrative staff allowed and expected to pick pens, the individual who took the last pen would be causally more relevant, regardless of her job status.

In summary, background assumptions against which the causal explanation is both asked for and given, can include normative and empirical expectations. As I argued above, descriptive rules can give rise to empirical expectations which take the form of a normative expectation: they are beliefs about how things should work, this is, what will happen unless something unexpected happens. I thus agree with Knobe[93] in that causal judgements include normative considerations, not only descriptive considerations. However, moral judgements are not the only kind of normative judgements. As Driver argues, "we are more likely to make attributions of causation to events that do not conform to norms"[94], but norms are not necessarily moral. Some norms may be merely statistical. Driver's position is that Knobe and Fraser's interpretation of their experimental results is too narrow, for it focuses only in the relevance of moral considerations, amongst other normative considerations. As Roxborough and Cumby have shown, a variation of the typicality of events affects their causal relevance. Their analysis supports Driver's more general claim: that norm violation or fulfilment is what makes some causal contributions more relevant than others. I agree in general with Driver's account; my aim here was to show that normative considerations serve indeed as reasons to highlight a causal factor rather than other because causal explanations are requested and offered against a set of background assumptions, which include both normative and empirical expectations, some of them being moral considerations.

Social commitments: the Queen of England problem

92 Schweder (1999: 117).
93 Knobe (2006).
94 Driver (2008: 459).

The Queen of England problem consists in the tendency to hold a promise-maker, rather than anyone else (such as Her Majesty) responsible for the consequences of failing to fulfil her promise (the death of a plant), by omitting to do what it took to fulfil it (such as watering that plant). As I discussed above concerning omissions, some authors argue that there is not a difference between the promise-maker and anyone else in what concerns their causal effectiveness. Nonetheless, the promise-maker plays a relevant role in the explanation of the death of the plant, while other actors do not[95].

The normality or frequency of "word-keeping" varies amongst pragmatic and cultural contexts[96]. In an abstract and ideal scenario, when agents engage in a promise, both of them are aware of the commitment, and word-keeping applies. This is the structure of the expectations that a promise generates: X (the promiser) promises Y (the promisee) to perform action Z, or to achieve state of affairs Z, giving rise to the following structure of interdependent beliefs:

- X and Y have the empirical expectation that X will Z, or Z will obtain
- X and Y have a normative belief: if X has a reason to Z (and none against), X should Z.
- Both X and Y believe that Z should obtain, in two senses:
  - It's what X ought to do, because he has promised to.
  - It's what will normally happen, because people is supposed to keep their promises.

A normative reason to Z – accepted by X – is also Y's explanatory reason for the normative fact "X should do Z", both in the sense of normality and in the sense of Z being mandatory for X.

Bringing back the Queen of England problem, Carolina Sartorio (in her own example) was under the obligation, which arose from a promise she made, to water a

95 See Beebee (2004); Braham and van Hees (2010).

96 Schlesinger (2008).

plant. She does not water it and the plant dies. The problem was that, as long as the Queen of England had not watered the plant either, she stands in the same causal relation to the death of the plant. The Queen of England problem is an example of what Barros calls "causal profligacy", or "causal promiscuity"[97]. Causal profligacy, or causal promiscuity, consists in considering every absence a possible cause of an outcome, as long as it is counterfactually true. For example, the fact that I am writing a Chapter of this Thesis counterfactually depends on not having been killed by Godzilla on my way to the library. But I would hardly say that I am writing this thesis *because* I have not been killed this morning. Contrastive models of explanation, Barros argues, successfully avoid the problem of causal profligacy, derived from accepting the causal effectiveness of omissions and absences.

An explanation of the death of the plant is usually required by those who expected the plant to be alive. For example, if Ann goes two months on vacation and she has not asked anyone to water her plants, it is quite likely that the plants will be dead by the time she comes back. The explanation of the death of the plants is Ann's going on vacation, because she was the only person expected to water them. If Carolina Sartorio has promised Ann to water her plants, then it is expected by Ann[98] that the plants are alive. Thus, Ann believed that:

- The plants will be alive, precisely because Carolina Sartorio is going to water them (and this is a reason to promise to water them in the first place).
- The plants should be alive if everything goes as it is supposed to go.
- Carolina Sartorio should (both it is normal to / has the obligation) water the plants, because she has promised to.

97 Barros (2011).

98 It is also expected by other persons who are aware of the promise and form the same expectations. If Bob is aware of the promise, he may wonder why the plants are dead, contrary to what he expected. But having the authority to demand or request an explanation requires to stand in a particular normative relation to the promiser.

When she comes back and finds the plants death, Ann is entitled to demand an explanation for this unexpected fact[99]. Given Ann's background beliefs, the Queen of England plays no role in the process that will keep the plants alive.

To sum up, the Queen of England problem can be solved as follows. It has to be shown why Carolina Sartorio, but not the Queen, is responsible for the plant's death (and not only exempt, but simply not responsible). The following explanation for the plant's death is offered: Carolina Sartorio's failure to water the plant, rather than the Queen's failure to water that plant, has caused the plant to die, rather than being alive. This explanation is counterfactually true, on the one hand, and explanatorily correct, on the other. Had Sartorio watered the plant, it would be alive. And the set of background assumptions, specially the expectations about what should normally happen (plus some background beliefs, such as "plants need water to survive"), and who is supposed to water the plant, make Sartorio's omission relevant, but not the Queen's. Hoekstra and Breuker provide a definition of the relation between causation and normativity: "negative causation occurs in the case where an actor should have performed some action, but didn't. It is therefore a normative statement, comparing perceived behaviour to some idealised standard behaviour"[100]. This "idealised standard behaviour" needs not to refer exclusively to moral standards. Judgements about what happens typically, or regularly, are also part of the standard against which an event or action is evaluated. Thus, causal relevance is related to the normative standard of behaviour expected from an agent. Promise making is a social practice whose goal is precisely to create such expectations; and therefore, promise-makers are more relevant in the explanation of the promised outcome, either its presence (by fulfilling the promise) or by absence (in the case the promise is unfulfilled).

99 This entitlement emerges both from the propositional and the action (social) commitment Sartorio has to Ann (see §4.2.1).

100 Hoekstra and Breuker (2007).

5.3.3. Explaining and justifying

Attributing responsibility is an evaluative process. Normative and empirical expectations serve to evaluate the role of the accused within the context of the production of the outcome. Excuses and exempting conditions modify the evaluation of the accused actions and motivational mechanisms. The difference between these two concepts is not clear in the philosophical literature. In the legal field, criminal defences are usually divided into justifications (self-defence, parental authority) and excuses (insanity, duress, mistake)[101]. In addition, some legal theorists distinguish between excuses such as insanity and excuses such as duress: while the former negate responsibility, the latter admit it. The first kind would represent an exempting condition, and the latter would be to offer an excuse for an action for which the agent admits responsibility[102].

In a philosophical context, Austin argues for the distinction between justifications and excuses: "In the one defence, briefly, we accept responsibility but deny that it was bad: in the other, we admit that it was bad but don't accept full, or even any, responsibility"[103]. By justifying our actions, we attack the claim that what we did was wrong. By excusing our behaviour, we explicit some conditions that palliate or mitigate our responsibility. Some excuses, such as insanity, exempt the agent from any responsibility. Exemptions would work as relievers of forward-looking responsibilities, or commitments, of agents. This is, those agents are not (at least normatively) expected to behave as it is expected from non-exempt agents. For example, a health problem can excuse an absence from work: the employee does something not permitted (not going to her workplace) because her health condition relieves her from that duty. Following Baron, "[e]xcusing someone in this sense amounts to exempting him from what would otherwise be a requirement, or at least an expectation"[104].

101 Robinson (1982); Berman (2003); Westen (2006).
102 Duff (2007).
103 Austin (1956: 2).
104 Baron (2006: 32).

I will suggest in this Section a different distinction, following what has been argued in this Chapter. In brief, I have argued that it is important to distinguish between responsibility as attributability and responsibility as accountability[105]. Both have a normative dimension: the former evaluates the agent's behaviour against a normative standard, given by the normative and empirical expectations of the attributer of responsibility. Those expectations, amongst other background beliefs, conform the context against which an explanation of the outcome is requested and given. Responsibility as accountability is an evaluation of the agent under the light of some social standards – moral or legal, for example. An agent cannot be accountable if it is not appropriate or correct to attribute her responsibility in the first sense. My suggestion is that, while exemptions affect attributability, excuses are offered in the accountability process[106]. I will argue in the first part of this Section that an agent is exempt when it is not possible to explain the outcome in terms of the agent's evaluation mechanisms, reasons or rational capacities. For instance, failing to fulfil the conditions for fitness to be held responsible, such as control or reason-responsiveness conditions, is a reason to be exempted from responsibility. In this sense, exemptions undermine the explanatory relevance of the agential capacities of the accused, because it traces the agent's action to an origin that undermines the explanatory relevance of the agent's authorship.

On the other hand, excuses have justificatory power. An excuse usually admits attributability, but mitigates accountability. Offering the reasons we had to act as we did can justify our actions. This is so by showing that what we did was actually normatively required by the reasons we had, and, if those reasons are evaluated by the attributer of

105 Watson (1996).

106 Wallace (1994) argues for a similar distinction between excuses and exemptions, although he uses those concepts with the opposite meanings that I defend here. Wallace argues that "excuses function by showing that an agent has not really done anything wrong" (Wallace 1994: 133). It would then be incorrect and unfair to hold the agent responsible. On the other hand, exemptions "identify the relevant conditions of accountability, explaining why it is unfair to hold people accountable when these conditions do not obtain" (Ibid., 154). Though I do not agree with Wallace's general approach to moral responsibility (in the Strawsonian tradition), his distinction between excuses and exemptions gathers an important point I will defend in this Chapter: the difference between considering that the agent is responsible, and judging that she deserves blame or praise (in the case of moral responsibility).

responsibility as good reasons, the behaviour of the agent is justified, and therefore not wrong. Of course, it is also possible to argue that an apparently praiseworthy action is not actually worth of praise, by stating the (wrong kind of) reasons the agent had to act. Also, excuses can offer a partial explanation, which does not relieve the agent from responsibility. They make the conduct more reasonable than before offering the excuse, but they do not make the action correct. This difference corresponds to the legal differentiation between justification and excuses, and it will be analysed in the second part of this Section.

Exemptions

If it is not correct or possible to explain the outcome in terms of what the agent did or omitted, then the agent is exempt. Of course, it has to be plausible that the agent indeed played some role; otherwise, we would all be exempt from shooting JFK, for various reasons (not being born yet, for instance). But we do not want to be exempt: we want to be simply not responsible, because we had nothing to do with the outcome.

Many cases of attributions of responsibility, specially when it comes to examine whether the agent is exempt, are problematic precisely because they support or contradict different background assumptions regarding the criteria an agent has to meet in order to be exempt. A typical case of controversy is drunkenness: a drunken person usually does not display a high level of self-control. She is not reason-responsive, in Fischer and Ravizza's terms (2000). Although she does not meet the conditions for being held responsible, it is usually considered that she is instead responsible for getting drunk in the first place. Sir Francis Bacon claimed that "if a mad man commit a felonie hee shall not lose his life for it, because his infirmity came by the Act of God; but if a drunken man commit a felonie, he shall not be excused because his imperfection came by his owne default"[107].

107 Bacon (1630: 34).

It is then possible to "trace" the agent's responsibility for an outcome to the initial conditions which lead to the outcome:

> Tracing is the idea that responsibility for some outcome need not be anchored in the agent or agent's action at the moment immediately prior to outcome, but rather at some suitable time prior to the moment of deliberation or action. [108]

It is possible to trace prior decisions and choices which lead to the actual actions, and it is also possible and common to anchor responsibility in the acquisition of dispositions or habits. For example, if an outcome is the result from making a decision motivated by greed, the agent can be held responsible for being greedy, in the sense that she is responsible for being greedy as well. However, the notion of tracing is problematic, because when we attribute responsibility for an outcome which depends on previous decisions or habits, we are supporting some claims about the voluntariness, control exerted and knowledge of the agent's actions, choices or habits.

To evaluate whether it is possible to trace the agent's decisions or habits to explain the actual outcome, and when anchoring the agent's responsibility to those previous states is valid, a normative standard of agent is needed, as I claimed in the beginning of this Chapter. Thus, attributions of responsibility require:

(i) that the agent meets certain control and reason-responsiveness conditions;

(ii) that the outcome is suitably explained in term's of the agent's decisions and actions (or omissions), so what is expected that the agent will do or decide serves as a contrast with what she actually did or decided;

(iii) that the agent can be held responsible for her own actions and decisions, so when tracing prior conditions, the agent can be held responsible for those conditions as well.

For (iii) to be attained, a normative account of agency is needed, such as the legal standard of reasonable agent. Attributions of responsibility are based on, and support, claims about this standard. For example, an agent should not drink if she knows that drinking will

108 Vargas (2005: 269).

impair her for doing what she should do; therefore, drinking does not make her exempt, despite her reasoning capabilities being undermined by the effects of alcohol. Tracing discloses some of this ideal agent's features, and reveals a set of background assumptions, specially concerning compulsion and perception of inevitability, that are problematic[109].

A typical case of non-exempting, but apparently inevitable behaviour is forgetting. Carolina Sartorio has promised to water Ann's plant; unfortunately, she forgets about her promise, and as a result, Ann's plant dies. Sartorio's forgetting can be hardly seen as voluntary: having made that promise simply did not come to her mind. However, depending on what normative standard of agent is being used, and Sartorio's actions compared to, it is possible to exempt (or at least excuse) Sartorio, while other standards would make her fully responsible. For example, a discussion about what level of foreseeability of forgetting to water the plant is minimally required may be necessary to settle the question. Scientific works on compulsion and addiction contribute to shape the standard of agent.

A similar but more common discussion has to do, in legal philosophy, with the concept of provocation. For example, an aggression can be justified (therefore excusing the agent) if the perpetrator has been provoked by the victim. In other cases, the perpetrator is exempt. Provocation can make an agent exempt if, as a result of the provocation, an agent suffers from a sudden and temporary loss of self-control, and the outcome of which the agent is exempt has been caused precisely by the agent's out-of-control actions[110]. I believe that provocation is a good example of how the causal relevance of the agent, and therefore the responsibility attributed to her, depend on both descriptive and normative considerations. If I walk down the street, accidentally step on a passer-by, and kill her in a fit of rage, it can hardly be argued that I had been provoked by the passer-by. Provocation depends on subjective and objective conditions. In the passer-by example, subjective conditions refer to whether the passer-by caused my loss of self-control, and whether my

109 For an analysis of four paradigmatic problematic examples, see Vargas (2005); and for a critical discussion, see Fischer and Tognazzini (2009).

110 Holton and Shute (2007).

killing her is the result of that loss. Even if those subjective conditions are met, for a provocation to take place, the objective condition requires that it is reasonable that I lose control because of the passer-by actions. At this point, a normative standard is needed. In the legal context, the figure of the "reasonable person" plays this role; but, of course, it is a changing standard, and highly context-sensitive. Most of the times, considering whether it is reasonable to act as someone else did, under specific circumstances, is not a straightforward inference. Besides, as I argued in Chapter 1, self-control is itself a controversial concept.

Collective, shared and mediated responsibility can also make specific agents exempt. Collective and shared responsibility will be analysed in Chapter 7. The former refers to responsibility attributed to a collective agent, while the latter is individual responsibility for the production of an outcome, which is the result of the sum of the individual contributions – for example, global warming. In those cases, the causal contribution of the agent is evaluated under the light of her role in the collective, or in the production of the outcome. Mediated responsibility, on the other hand, consists in tracing responsibility to another agent's decisions and actions[111]. For example, military chains of command "pass the buck" towards higher ranks. However, mediated responsibility can take more subtle forms. For example, by providing certain incentives, it is possible to manipulate the other's behaviour (see Chapter 3). Most of the times mediated responsibility can excuse an agent and make another responsible; but the mediation needs to be strong enough to explain the outcome without appealing to the exempt agent's actions.

Finally, it is important to clarify the relation between exemptions and the criteria an agent has to meet in order to be held responsible (§5.2.1). If the agent does not meet the relevant control and reason-responsiveness conditions for being held responsible, but she plays a role in the causation of the outcome, she is usually exempt. The lack of those conditions makes it incorrect to explain the outcome in terms of her agency. In the

111 Attfield (2009).

example of provocation, the agent is not being reason-responsive, and she lacks self-control. In other cases, the agent might lack some relevant information, which could not have been obtained, and therefore she meets the control conditions, although she cannot respond to the relevant reasons to choose or act accordingly. However, I think there are cases in which the agent meets the criteria above, and nonetheless it is not correct to explain the outcome in terms of what the agent did or chose. Therefore, not all the exemptions require that the agent does not meet any of the conditions for fitness to be held responsible, in Pettit's terms.

To illustrate this possibility, let's recall Mr Puppet. He is not under the direct control of the group of scientists; instead, they have set up all the variables (with a margin of error equal to 5%) for shaping him: his character values, beliefs and intentions. The actual outcome is that Mr Puppet has committed murder in a drug deal gone bad. Mr Muppet does not lack any control or reason-responsiveness conditions. However, I believe there is a difference in the explanation of the murder depending on whether Mr Muppet's committing that murder was or was not a part of the scientists' plan. Was it part of the plan, the murder would be more easily explained by appealing to the group of scientists' plan, rather than Mr Muppet's motives to commit the murder; in that case, the explanation of his motives would all point to the scientists' plan. On the other hand, if Mr Muppet acted unexpectedly, thus changing the planned course of action, he can be held responsible for the murder. Or course, the group of scientists would also bear some responsibility, for they have set up the conditions that made it more likely that Mr Muppet committed murder, and the explanation of the murder will also include the scientists' plan.

To conclude, the reasons why an agent is considered exempt are the reasons why this agent's causal contribution is not explanatory of the outcome. Usually, if an agent does not fulfil the control and reason-responsiveness conditions, she is exempt. But there are situations in which the agent does fulfil these criteria, and is nonetheless exempt. This is so because the reasons for considering a contribution explanatorily relevant are context dependent.

Excuses and justifications

Offering an excuse consists in arguing that, although responsible in the sense of attributable, the agent is not accountable for what she has done. In this sense, offering an excuse is similar to justifying one's conduct, to offer some reasons that the attributer of responsibility should (from the point of view of the accused) take into account when evaluating the conduct of the agent and the outcome for which the agent is responsible. Berman[112] argues that the difference between justification and excuses lies in the distinction between acting wrong and deserving blame (or punishment). From a legal point of view, a justified action is not criminal, whereas an excused agent has committed a criminal act, but is not punishable. From the moral point of view, a justified action is not wrongful, while an excuse is offered when the action is wrongful but the agent is not blameworthy[113].

Both justifications and excuses have to do with reasons: they consist in presenting the reasons the agent had to act as she did. These reasons are explanatory (or motivational), and they can also be normative. Duff[114] has discussed the relation between reasons and excuses, and the justificatory force of reasons used as excuses:

> To excuse my action is to admit I had conclusive reason not to act as I did – that I acted either against a categorical, indefeasible reason, or against the balance of reasons; but to plead that I could not reasonably have been expected to act in accordance with either that categorical reason or the balance of reasons – which is to say, since the expectation that is involved here is clearly a normative one, that I cannot reasonably be condemned for failing to act thus. To offer an excuse is thus to admit responsibility, but deny liability: I admit to committing an action for which I must now answer, but seek to block the otherwise legitimate transition from responsibility to liability (liability, in this context, to moral

112 Berman (2003)

113 I sympathise, in general, with Berman's account; I do not agree, however, in including insanity, infancy and involuntary intoxication as a kind of excuse, rather than an exemption. Following his account, a two-years-old child who stabs another child has done something wrong, but is not blameworthy. Although I believe that stabbing people is wrong, I do not believe that the toddler has done anything wrong. I would not believe either that a dog biting someone is doing something wrong, although I believe that a normal person biting someone (let's say, just for fun) is doing something wrong. The wrongness of the action does not only depend on the evaluation of the outcome, but also an evaluation of the agent; a toddler can be hardly classified as a full-capable agent.

114 Duff (2007); (2009b).

criticism or blame) by offering an exculpatory answer. [115]

Duff argues that justifying one's actions serves to block the transition between responsibility and liability. In Watson's terminology, this would entail to block the transition between attributability and accountability. This is, the agent admits her authorship, but presents an argument against the evaluation of what she did as wrong (either form a moral or a legal point of view).

The debate on whether excuses admit responsibility (Berman) or negate it (Austin) lies in different conceptions of what an excuse is, but also in different concepts of responsibility. Similar to Watson's distinction, Duff argues that responsibility as answerability is different from responsibility as liability. Excuses affect liability: the practice of excusing entails that there is something to excuse. If I have been pushed, and as a result, a vase is broken, I would not say "Yes, I broke the vase, but I did so because I was thrown at it"[116]. Breaking the vase is not something I did as an agent, and it would be more correct to state that "the vase broke because Peter threw Miranda at it". In this case, I would be exempt: I have nothing to excuse, for I "did" nothing. Peter pushed me just to bother me; and he is responsible for the breakage of the vase. Now, let's imagine that Peter pushed me because he thought I would just sway in a funny way, then he could offer an excuse: "Yes, I pushed Miranda, but I did not know that my pushing would make her fall". He can claim that my fall was an accident, because both my fall and its consequences had not been foreseen. The force of the excuse will be related with the expectations of the attributer: it would be reasonable to foresee that there is a vase near me in a vase shop. In that case, Peter would have acted negligently, and his excuse would not have justificatory force. However, offering excuses serves also to clarify the agent's intentions in order to mitigate accountability. For example, Peter's excuse is justificatory had he been accused of making me fall on purpose.

Thus, I will consider that excuses admit responsibility, but have the capacity to justify why one acted as she did. Justification, as well as certainty, is not a binary concept,

115 Duff (2007: 53).

116 This example is a simplified version of the window example presented in Duff (2009b: 303).

but it is instead gradual. Stronger evidence makes us more certain that something is the case; stronger reasons makes our actions more justified. The difference between justifications and excuses as described in the legal literature is analogous to the strongest and less strong justifications. In the strongest case, the agent not only accepts responsibility for what she has done, but offers an argument showing why she should have done precisely that. A complete justification, I will claim, is neither an excuse or an exemption: it is an argument against what the agent is being attributed responsibility for. Excuses, on the other hand, provide pro-tanto reasons, in Broome's terminology [117]. They might not completely justify what the agent has done, but has (at least some) justificatory power. The reasons provided by excuses are explanatory and justificatory reasons[118]. For example, Peter offered as an excuse that the reason why he pushed me was that (he thought) I was going to sway in a funny way. Swaying in a funny way does not justify to push someone, but it explains why Peter pushed me. His excuse partially justifies what he did: he did not know that I was going to fall (let's suppose it is reasonable to believe so), nor that there was a vase near me. Excuses, then, have the purpose of influence responsibility as accountability.

Justifying an action has the same purpose. For example, a doctor who is administering a patient morphine to relief her from her suffering knows that her patient has a high probability to die as a consequence of the drug. When attributed responsibility for the death of the patient, he can argue that he had the permission (due to medical authority and the patient's will) to do it, and that it is not wrong to administer morphine to terminal patients. Thus, the doctor is justified in administering morphine, and responsible, in the sense of attributability, for the patient's death, because it was a foreseen risk of administering the drug. However, the doctor is not accountable: she did what she ought to do[119].

117 Broome (1999).

118 The difference between offering an explanation and offering a justification is not clear cut; as shown in Chapter 2, normative and explanatory reasons can overlap.

119 I am aware that this example is highly controversial. I assume here that the medical procedures for terminal patients are morally correct.

In this sense, excuses acknowledge that the agent has violated some normative or empirical expectations, but that there is an explanation for that violation. A good excuse usually explains and (partially) justifies the agent's actions in the context of what is expected from her. A bad excuse can be fully explanatory, but does not justify the action. Explanatory or motivational reasons cannot be good or bad reasons (although they can be strong or weak): they just state what motivates an agent. Normative or justificatory reasons, on the other hand, can indeed be good or bad reasons, depending on the strength of their justificatory power. On the other hand, justifications argue against the normative (and, in some cases, empirical) expectations that the attributer of responsibility has. In the example above, justifying to administer morphine to a terminal patient implies to claim that it is normatively expected from the doctor to do so – perhaps through arguing the normative reasons to guarantee a terminal patient a painless death. Thus, justifying an action does not deny responsibility in the sense of attributability: as I argued before, responsibility in this sense does not require a moral evaluation of the agent.

Regarding violations of social commitments, justifications would consist in showing that there was an overwhelming reason (or more than one) to violate the commitment. Bringing back the Queen of England scenario, Carolina Sartorio had promised Ann to water her plant while she was on vacation. Sartorio fails to do so, and the plant dies. Sartorio might offer the following justification: “I know I made a promise and failed to fulfil it. However, I came to know that Ann had three cats. I am extremely allergic to cats; if I am exposed to one, I would spend several days in the hospital. I tried to find someone else to water the plant, but I found nobody. Thus, I decided not to water the plant”. While it is not denied that promises should be kept, and it is correct that people is expected to keep their promises, it is claimed that avoiding a serious health problem is a strong enough reason to let a plant die. Thus, Sartorio admits responsibility: the plant died because Sartorio did not water it; but she did not so because of a good reason. It is not expected from allergic people to expose themselves without a good reason to do so.

Excuses, as argued above, also have justificatory power, although they do not provide a full justification. When accused of causing the plant's death, Sartorio may offer the following excuse: “I know I broke my promise, but I heard on the radio that it was likely to rain those days. I then went on holiday myself without looking for nobody else to water the plant, assuming that it would be probably naturally watered”. In this case, it is not expected from her, as in the case of justifications, that she does not water the plant. Having heard a weather forecast on the radio is not a good evidence for making sure that the plant will be watered. However, Sartorio's excuse explains why she unfulfilled her promise and, had it rained, her promise would be probably considered fulfilled. After all, the promise consists in making sure that the plant does not die (through watering it); but achieving this goal through other means is not necessarily prohibited.

Lastly, Sartorio can also argue that she is exempt from any responsibility. This would entail that the explanatory claim “the plant died because of Sartorio's failure to water it” is incorrect. For example, she can argue that she made the promise whilst she was drunk. Ann knew that Sartorio was drunk, and nonetheless made her promise that she would water her plant. Sartorio may argue that this was not a real promise (see §4.2). It was reasonable to think that Sartorio would have forgotten, by the next day, the promise she had made. Thus, Ann's empirical and normative expectations about Sartorio watering the plant were not reasonable, and they cannot serve as background assumptions against which to demand an explanation to Sartorio about why the plant has died.

PART III: INDIVIDUAL, SOCIAL AND COLLECTIVE COMMITMENTS

In the previous Chapters, I have shown that the normative features of social commitments can be explained through the normative requirements of rationality, and the exercise of the agent's normative powers. I have shown that agents are able to socially adopt normative reasons for action, and to confer rational authority over these reasons to others. By engaging in a social commitment, an agent accepts a social interaction (a request, a promise, a command, an agreement...) as having normative force over her actions. An agent has rational authority over her reasons: she can change her mind, and change her normative judgements in the light of new normative reasons. This is why self-directed commands are suspicious: the agent can, in principle, revoke them at will[1]. However, when an agent becomes socially committed to another, she accepts that the creditor, to whom the commitment is made, has now the capacity to judge that the debtor ought to fulfil her commitment. This is why a commitment which is violated for very good and morally praiseworthy reasons is a *violated* commitment nonetheless.

In Chapter 4, I explained that changing one's mind is a way of *exiting from* the commitment. The normative requirements no longer apply, insofar we have changed our normative judgements, and maybe also our intentions. Thus, we can release ourselves from the obligation to comply with those requirements—at least, those we incur because of having the reasons or intentions we have abandoned. Social commitment differ in the sense from practical commitments: a promise made is a debt unpaid, unless the debtor is left of the hook by the creditor.

1 As I showed in Chapter 2, an agent cannot change her normative reasons at will: she must have other normative reasons to do so. However, the agent is normatively *autonomous*.

My aim now is to show that collective commitments can be analysed using these same theoretical tools. Collective agents can adopt a wide variety of forms and structures. Certain institutionalized collective agents, such as companies, universities and governments, require a specific and well defined kind of membership, and the roles within the group are explicit. Loosely structured groups, such as spontaneous basketball teams made up by individuals who happened to meet by chance in a basketball court, have a quite different normative organization. Thus, given the wide scope of the collective agency phenomena, my aim is not to provide an exhaustive account of the different kinds of commitments that can be found in different collective entities. I will focus instead on two general kinds of commitment: those involved in membership as *affiliation*, and those acquired by a collective agent, similar to practical individual commitments.

On the other hand, I have shown in Chapter 5 that the fact that an agent is under some normative constraints, which are imposed (amongst other factors) by the commitments acquired by that agent, affects the way this agent can be attributed responsibility for certain outcomes. Particularly, the agent's commitments, both individual and social, can explain why a particular omission, amongst all the things al agent fails to do, can be properly attributed to an agent. An outcome can be explained through an agent's omissions when that agent stands in a particular normative relation with those omissions, such as being socially committed to the performance of the act omitted. This framework can also be applied to collective agents, in cases in which the outcome is explainable through collective actions or omissions.

The following two Chapters are structured as follows. Chapter 6 is devoted to the analysis of collective commitments. In order to distinguish mere aggregations of individuals from collective agents, it is necessary to appeal to a distinct kind of membership that is necessary to make up a collective agent (§6.1). Membership as affiliation, as opposed to membership as aggregation, is a kind of social commitment between an agent and a group, or between two or more individual agents who, through their social commitment to become members, create a collective agent (§6.1.1). The fact

that becoming a member entails to engage in a social commitment explains the normative structure of the member's obligations which, from the point of view of rational requirements, are the obligations that a debtor (the member) has towards the creditor (the group, to which the member belongs, *and* the other members). Ceasing to be a member constitutes the same process as asking for release from a social commitment, described in §4.2.2. The difference lies in that, in the case of membership, the member is also part of the collective agent, so she can take part in the decision process (§6.1.2.). Becoming a member, then, requires to accept the group's practical commitments as normative reasons for action, over which the member shares her rational authority with the other members, through an exercise of her normative powers. Collective practical commitments are analogous to individual practical commitments (§6.2). They are subject to the rational requirements explained in §2.2: *enkrasia* and *resolve*. Thus, the collective agent incurs in a rational obligation when getting committed to a goal. The discursive dilemma seems to pose a challenge to the consistency between the member's and the group's judgements, intentions and goals. This inconsistency shows that the decision procedure the group adopts also serves as a normative reason to accept the conclusion.

The aim of Chapter 7 is to apply the framework of responsibility attributions developed in Chapter 5 to collective agents. I will deal in this Chapter with two kinds of aspects of collective responsibility. First, I will argue that responsibility cannot be attributed to a group of individuals who fail to conform a collective agent (§7.1.1). This would be a case of *shared individual responsibility*; collective responsibility requires a collective *agent* to be attributed. Thus, the collective agent must fulfil certain agential conditions that makes it suitable for being the target of responsibility attributions (§7.1.2). While a collective agent is the necessary target of collective responsibility, a collective action does not need to have taken place: a collective omission, I will argue, can be the basis for attributing collective responsibility. Second, the problem of how collective responsibility should be distributed will be addressed (§7.2). I will present three confronting perspectives on whether collective responsibility is shared amongst the

members, and if so, whether it is *diluted*: the greater the group, the lesser the responsibility each member holds (§7.2.1). I will argue that insofar membership as affiliation requires acceptance, individual members always share collective responsibility: this is why mere bystanders can be individually, but not collectively, responsible for failing to act jointly. The degrees of responsibility will depend on the agent's role in the decision processes and in the actual production of the outcome: this is why the different roles an agent can have within a group are determinant. Finally, in §7.2.2, I will bring back the discursive dilemma, and consider whether inconsistencies between the individual and the collective level give rise to responsibility voids, i.e cases in which the collective agent is responsible, but its members are exempt. I will argue that the discursive dilemma does not show that there are responsibility voids, because even in cases in which the members do not personally believe that the group ought to achieve a goal, they have accepted this normative judgement nonetheless, and thus they are committed to its promotion as members, which eliminates the possibility of exemption.

Chapter 6. Collective Commitments

Collective agents are social in nature; nonetheless they exceed the scope of social commitments. This is so because collectives, as agents, are capable of acquiring practical commitments, in a way that is similar to individual practical commitments. A group is conformed by a set of social commitments amongst members who, at the same time, are also supposed to be internally committed, as I argued in §4.2.2. However, groups also have the capacity to acquire further practical commitments, towards individual agents, or towards other groups. Binding ourselves to other agents, either individually or collectively, and constraining our actions because of having acquired those commitments is a distinctive capacity of (rational) agents, either individual or collective.

In this Chapter, my aim is to show that collective commitments require membership, on the one hand, and the capacity to evaluate and accept as valid reasons for action, normative judgements and goals, on the other.

6.1. Membership as a social commitment

The scope of collective agency is so broad that it hardly refers to a unified concept or entity[1]. In fact, the same goes with the concept of agency; in this dissertation, I have

1 Just to mention two different approaches (although there are many more), it can be argued that a group is defined by holding intentions which are distinct and independent from individual intentions:
That is, a number of individual human beings form a collectivity if and only if:
(i) they act in ways whose significance can be adequately captured only by an ineliminable reference to some corporate body as part of which they are acting, where
(ii) what that corporate body does is distinct from anything which they as individuals do, and where
(iii) the corporate body is a persisting one whose survival is relatively indifferent to the persistence of the particular individuals which compose it at any particular moment. Graham (2002)

focused on rational agency, particularly, its relation with deliberation and rational requirements. Even if I aimed to perform the same task applied to collective agents, I would still face a difficulty: the great variety of collective agents and actions. In its simplest form, a collective agent requires two individual agents who engage in performing an action together, not merely next to each other, but as an action that can be attributed to them as a group. Examples of collective agency include painting a house together[2], walking together[3], or playing a football game[4]. Similarly, the Catholic Church, a football team, a corporation, the Communist Party of Ruritania[5] or Britain and Argentina[6] are collective agents. Furthermore, any random collection of individuals, under the appropriate circumstances, are able to form a group agent. Thus, collective agency and actions cover a wide spectrum—to wide, I believe, to form a unique kind of agent with a certain normative structure. I do not aim here to provide a theory of the normative bonds of and within all kinds of collective agents, between what constitutes an agreement between you and me to paint a house together, and an agreement between British Airways and the Spanish government in order to buy (and sell) the Iberia company, or the signing

---

Alternatively, the role of members can be stressed as a necessary condition:

> A collective g consisting of some persons (or in the normatively structured case, position-holders) is a (core) we-mode social group if and only if:
> (1) g has accepted a certain ethos, E, as a group for itself and is committed to it. On the level of its members, this entails that at least a substantial number of the members of g have as group members (thus in a broad sense as position-holders in g) collectively accepted E as g's (namely, their group's, "our") ethos and hence are collectively committed to it, with the understanding that the ethos is to function as providing authoritative reasons for thinking and acting qua a group member;
> (2) every member of g as a group member "group-socially" ought to accept E (and accordingly to be committed to it as a group member), at least in part because the group has accepted E as its ethos;
> (3) it is a mutual belief in the group that (1) and (2). Tuomela (2007: 19–20)

I will here suggest that collective agents admit great variation, and thus some can survive its members, while others do not; and some require that its members explicitly accept the group's commitments, while others allow for implicit acceptance.

2 This example is from Bratman (2009b).

3 This is the paradigmatic example in Gilbert (1990).

4 See French (1998).

5 See Tuomela (2007).

6 See Copp (2006).

and ratifying of the Kyoto Protocol by 191 countries. This task would largely exceed the scope of this work.

Collective entities allow then for a great variety of normative configurations, this is, who should do what, as a member of the collective. Certain groups require that the agents are aware and conscious about their belonging to the group, while other groups allow for looser configurations, in which agents engage in social activities without being aware that they are in fact part of a collective agent. Furthermore there are some outcomes that are the product of an aggregation of individual actions, and so they are "collectively" produced—for instance, the case of pollution. Cases of overdetermination (see §5.2.2), in which an individual contribution does not have a significant causal effect on the outcome, are paradigmatic of collective activity. Thus, we find collective action without a collective agent. The opposite is also possible: to have a collective agent who fails to perform an action. The agent can be nonetheless responsible for the outcome if it is appropriately explained by the group's omission, and yet we have a collective agent that has not performed any (collective or non collective) action.

Collective agents are formed by individual agents, who stand in a normative relation with the group, as well as with other individuals within the group. In this sense, the concept of group is broader than the concept of collective agent. A group can be made up individuals who share a common feature: a school class, a group of the hundred richest persons in the world, the group of those affected by certain disease. The members of these groups are members by *aggregation*: they share a common feature, but their membership does not turn the group into a collective agent. On the other hand, certain groups require membership as *affiliation*: the individuals must do something, or omit doing something, in order to become members of the group. Groups made up by aggregation of individuals may be subject to *restricted general obligations*[7]. When the norm "cyclists must give way to motor vehicles" applies, for every agent who is a cyclist, the obligation holds. Restricted general obligations are always imposed, for the group lacks of a decision mechanism in

7 Royakkers and Dignum (1998).

order to voluntarily acquire al obligation towards another agent, or a practical internal commitment. Restricted general obligations thus differ from *collective obligations*, those whose target is a collective agent. Those obligations can be voluntarily acquired (either through an internal or a social commitment), or socially imposed (through attributing the obligation to certain collective agent). For example, "corporations ought to pay taxes" is a collective obligation attributed to corporations, which are collective agents[8].

Given that membership as aggregation does not qualify groups as collective agents, I will use the concept of membership as referring only to affiliation hereafter. The concept of membership, then, refers to a specific relation, holding between an agent and a group, which can be conceptualized as a social commitment between agents (either individual or collective) to promote a collective goal, or a set of goals. This relation does not exist in virtue of some of the agent's *features*; rather, it is the result of the agent's *actions*.

It is frequently claimed in the literature on collective agents that group members need to hold (or to share) a specific intentional state, or to satisfy certain epistemic conditions[9]. For example, Searle[10] claims that each of the members has to hold a *we-intentio*n: the (individual) intention that *we* do φ. This is, each individual has a mental state whose content is a plural subject. Bratman, on the other hand, claims that what is distinct of *shared intentions* is not their content (as Searle states), but their attitude. He suggests that the following conditions are involved in a shared intention:

(i) intentions on the part of each in favor of the joint activity,
(ii) interlocking intentions,
(iii) intentions in favor of meshing sub-plans,
(iv) beliefs about the joint efficacy (in conformity with the connection condition) of the relevant intentions,
(v) beliefs about interpersonal intention-interdependence,
(DEP) interpersonal intention-interdependence, and
(vi) common knowledge of (i)–(vi) and (DEP). [11]

8 In this example, the norm "corporations ought to pay taxes" is also a restricted general obligation, applying to the group formed by all the existing corporations, which is a group by aggregation, but not a collective agent.

9 See Chant and Ernst (2008).

10 Searle (1990).

11 Bratman (2009b: 54). This is one of Bratman's latest formulations; but see also Bratman (1992); (1993).

The analysis of the interdependency of intentional states is useful to assess the problem of how joint or shared intentions are created. However, commitments to belong to a group, or the commitments a group agent can incur on, exceed the scope of interdependent intentions. A dishonest commitment is still a valid commitment, and it does not entail that the agent holds certain intentions, but she is *supposed* or *attributed* those intentions. Thus, the set of interdependent intentions mentioned above are *normatively* (and usually *empirically*) *expected* from the agents involved, although they are not necessarily present in order to *create* a collective agent, capable of acquiring internal and social commitments. Analogous to social commitments, collective commitments prescribe, but do not entail, mental states. Of course, if that agent is willing to fulfil its commitments, individual intentions are indeed required.

Thus, in order to create a collective agent, the involved individuals must become socially committed amongst them, being the object of the commitment that they will acquire a certain (practical internal) commitment to do something, either collectively or individually. In this sense, collective agents are created through membership as affiliation, in a way that is similar to synallagmatic agreements, this is, to agreements that give rise to reciprocal obligations.

6.1.1. Becoming a member

Social commitments involved in the adoption of a collective commitment, and thus constituting a collective agent, have a similar structure to social commitments between individual agents. Instead of conferring the creditor the right to release the debtor from her obligations, the two agents[12] involved confer this right, through the exercise of their normative powers, to themselves as a group. This means that, in order to be released, they have to reach an agreement about considering that other reasons justify that they

12 Of course, more than two individual agents can compose a collective agent. I will focus on groups made up by two agents as the simplest form of collective agent.

ought not to do what they have agreed to do. A unilateral judgement violates the collective commitment—perhaps for good reasons, but violates it nonetheless.

The goal to which A and B are collectively committed is a collective goal: the goal that A and B do φ. From the perspective of the agents, this goal involves a collective agent to which they belong. Thus, A and B share the goal that "we do φ". By agreeing that they will do φ, the also acknowledge that having committed themselves to φ is a reason to do φ. In this sense, collective commitments are social commitments, in which new reasons for action are created. Practical individual commitments, as I explained in §2.1.3, do not generate new reasons for action, as this would be a case of bootstrapping. In order to commit oneself to do something, an agent changes her attitude towards the chosen goal, but does not have new facts available that serve as reasons to do what the agent is committed to. Social commitments, on the contrary, are *actions*, which serve as reasons that justify the content of the social commitment, in particular, *exclusionary* and *second-order* reasons to do it (see §4.2.2).

Hence, both agents agree that they ought to do φ, because of their commitment to do φ. A and B are, individually, debtors; and A and B, as a group, are the creditor. This means that the can release themselves from the commitment, but they have to do it collectively, in the same way they have created the commitment. In this sense, membership is a kind of social commitment to create a practical collective commitment. My claim is similar to Gilbert's theory of joint commitment. Gilbert has extensively argued[13] that joint action has a normative (and non-moral) dimension. A joint commitment requires that each individual participant express her readiness to participate in the collective task, conditionally to the other agent's readiness to do the same[14]. When all participants jointly express that they are conditionally committed to do something collectively, i.e. as a body[15], this commitment becomes unconditional for them as a group.

13 See Gilbert (1990); (1992); (1999); (2006); (2009).
14 This definition is from Gilbert (2006: 140).
15 Gilbert (2009: 179).

Gilbert takes the concept of joint commitment to be a primitive concept, required for sharing an intention to act collectively.

In outline, I agree with Gilbert's account in that a collective agent holding a goal is composed by individual agents jointly committed. I endorse the claim that collective agent is intrinsically normative, although not necessarily moral. However, I believe Gilbert fails in distinguishing the kind of commitment involved in membership, which is a reciprocal social commitment amongst individual agents, and the kind of commitment the collective agent adopts[16]. She takes both to be part of joint commitment, which is a commitment to act as a body. The subject of this commitment are both the individuals and the collective agent. It is common to overlap these commitments, insofar membership is a social commitment to adopt the collective goal, thus generating a collective commitment when all members agree to. However, I believe it is appropriate to analyse them separately, dues to the fact that the normative requirements applying to each kind of commitment are different. A collective can fulfil its practical commitments while having one member who has abandoned the collective (and therefore has violated her social commitment with the other members); and a collective can also be akratic or weak-willed if its members revise its goal and contradict a previous practical commitment of the agent, while no individual member has individually violated her social commitment.

In fact, Gilbert's concept of joint commitment is useful to analyse simple forms of collective action. But when the collective agent is more complex, a distinction between the normative requirements of membership and the normative requirements of the collective agent (derived from its practical commitments) becomes necessary. Gilbert's example of two persons walking together is a collective agent in the simplest form: its members are socially committed to do (between them) what the collective agent is internally committed to do: walk together as a body, in Gilbert's terms. However, suppose that I get a job as a postdoctoral research fellow at the University of Ruritania. I then become a member of this institution, but not by committing myself to the other members to achieve goals φ, χ

16 In fact, the notion of collective or joint commitment often confuses different levels of individual commitment; see Bouvier (2007).

and ψ, which are the goals that this institution is trying to achieve at the moment of my incorporation to the group. Rather, I commit myself to the collective agent to do my part in the production of the University intended outcomes. But I do not commit myself to perform the University's goals: my social commitment differs from the University's internal commitment. Castelfranchi[17] argues that the kind of social commitment required to belong to a collective is a *generic meta-commitment*. It is generic, on the one hand, because, following the example above, I commit myself not to do a specific action, but to do any action from a set of actions, which are usually part of the plan that the collective agent has to achieve its internal practical commitments. On the other hand, it is a meta-commitment: it is a commitment to commit myself in the future, when it is required by the collective agent—the University, in this case. Castelfranchi labels this kind of social commitment *organizational* commitment:

> When the member x of the group X is organizationally Committed to his group, he is Committed to accept the requests of the group within a certain class of actions (his office). Then, x's Org-Commitment to X implies that if there is a request of X to x about an action of the class A, x is automatically S-Committed to X to this action, he automatically gets an obligation to do a. [18]

To sum up, the way an individual agent becomes a member of a collective agent depends on the complexity of that collective agent. In the simplest case, two individual agents can socially commit themselves to each other to do something together, as a group—thus, engaging in a reciprocal commitment. However, complex collective agents, such as companies or institutions, continue to exist after the initial members are gone, through the inclusion of new members. In this case, the process through which an individual becomes a member consists in acquiring a social commitment with the collective agent, by means of an agreement between certain operative members. For example, in order to be hired by the University of Ruritania as a postdoctoral fellow, I sign a contract whose parties are the University and myself. Certain individual agents whose role allows them to

17 Castelfranchi (1995).

18 Ibid., 47

represent the collective agent sign the contract on behalf of the University; and the University and I become reciprocally committed to each other. As a members, I acquire general obligations or meta-commitments, and also some rights conferred by the University—for instance, the right to get paid every month. Similarly, the University has acquired some rights over me—that I teach some undergraduate courses, or collaborate with an ongoing research, for example—and also certain obligations, such as paying my salary, or allowing me to park my car in the University parking space.

6.1.2. Exiting from membership

I argued in §4.2.2 that the conditions under which an agent can exit from a normative requirement (and thus abandon a commitment) can vary, depending on whether the agent has acquired a practical individual commitment, or a social commitment with another agent. In the former case, reconsidering one's reasons, normative judgements or intentions is a legit way to exit from the requirement. On the contrary, if the agent is socially committed to another agent, she has to be released by this agent. Of course, she can violate the social commitment (for example, a promise) for sound and good reasons, but she is violating the commitment nonetheless. In the case of the social commitments involved in membership, the scenario is more complex.

If the group in question is an instance of the simplest case discussed above, this is, two agents engaging in a social commitment to achieve a collective goal, then the picture is analogous to the requirements of social commitments (see Figure 4). None of the agents can unilaterally cease to be a member, thus exiting from the normative requirement. In this case, the collective has to release the member, because no member has the right to unilaterally rescind the obligation acquired[19]. As we can see in the Figure below, the social commitments of which membership consists are identical to the internal practical commitment of a collective agent. Thus, if that collective agent, or any of its members, wants to exit from its practical commitment, it needs to be released by the same collective

19 Gilbert (2006: chap. 7) also stresses this point as a central feature of joint commitments.

agent. As I stated above, in a collective agent consisting of two members, A and B, A and B are, separately, debtors; and A and B are, jointly, the creditor.

In a social commitment between individual agents, such as a request, the creditor and the debtor are different agents; in the case of collective commitments in simple groups, both agents belong to the collective agent and thus have some normative authority over the group's commitments, although this authority is shared. To put it differently, it is like sharing the ownership of a good, for instance, a house. Both agents have the right of usufruct, and they have to decide together whether to sell the house. Or course, dissent is possible, and frequent. If A and B cannot agree on what to do, what commitments to accept, or whether to revoke its commitments, the group is quite likely to disappear.

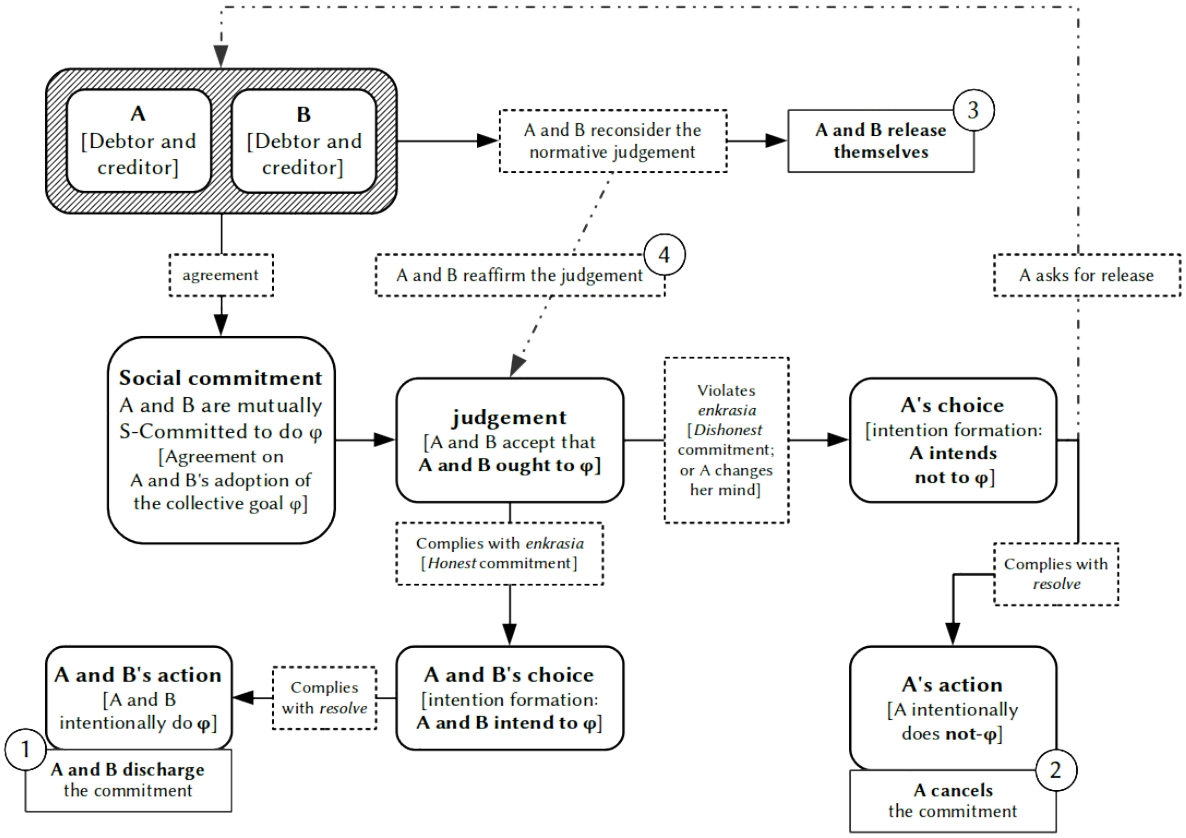

Figure 4: Membership as a social commitment

A and B agree (through a social commitment) to acquire a collective commitment: the agree that they are going to do φ together. Through this agreement, they become members of the collective. They accept, then, that they ought to do φ in virtue of their commitment to do φ, just as it happens with social commitments: an agent ought to do what she has promised in virtue of her promise. In the Figure, then, A intends not to φ. She can either do not-φ intentionally (complying with the *resolve* normative requirement), but that option will violate the commitment, being unilaterally cancelled (number 2 in the Figure). A can alternatively ask for, suggest, or request[20] a release. This option goes back to A and B, as a group, and they reconsider their commitment. They can revoke the commitment, which would release not only A but also B from her obligations (number 3); or they can judge that the commitment stands (number 4), and A and B keep their individual commitments as members. In case A and B do not reach consent, the group is likely to disappear, mostly because the lack of consent indicates that one of the members of the group is not willing to continue belonging to the group; in this case, if this member cannot persuade the rest of the members to release her, she will go probably cancel her commitment anyway, depending on the costs of doing so. In the same sense that an agent can always break a promise—i.e. it is something that the agent is able to do, although she is incurring in the violation of a normative requirement, a member can always abandon the group. This is why groups by aggregation do not have a strong concept of membership; for example, there is no way to exit from the group formed by all human beings.

There is a possibility concerning membership as affiliation that seems to undermine the claim that membership is a social commitment. Each group has the authority to decide the conditions under which an agent can exit from the collective in a legit way, this is, without cancelling her commitment. A group can be such that it decides that a member can cease to be a member, any time she wants to, without asking for release, and her commitment will not be violated: she can rescind her own membership.

20 This depends on how the group has agreed to proceed if a members wishes to revoke the commitment.

Imagine that a member releases herself by declaring: "all right, people, I have been collaborating with you so far but it is time for me to leave; good bye and good luck with your goal". I believe that, in cases like this, the weight of each members reasons to abandon the group are jointly accepted. Its structure is then similar to the social commitments made to Merciful Merle from §4.2.2. Any time a member finds a reason that makes her judge that she ought to abandon the group, if she communicates that reason to a group that reacts as Merciful Merle, the group will release her from the social commitment she acquired.

To put it differently: when an individual becomes a member of a collective agent, being part of the collective is a reason for that agent to promote the group's goal. The reason the agent has because of belonging to a group is content-independent, like those created in a standard social commitment. This is, they do not directly count if favour of doing what that specific group is trying to achieve, but in favour of doing whatsoever the group tries to achieve. They are also exclusionary, because they affect other reasons: they are second-order reasons (see §4.1). It is in this sense that membership is normative in the first place although, of course, moral and legal norms may apply as well. As Schmid points out, the debate between normativist and non-normativist accounts of collective action usually miss this kind of normativity:

> To the individuals involved, a collective intention provides a reason to form an appropriate personal intention (i.e. the intention to perform one's part, or to we-intend the collective activity). Contrary to what the existing normativist accounts suggest, the basic sense in which one ought to (intend to) do one's part is not that of social normativity (duty or obligation to (we-)intend x), but of pre-social normativity (having a reason to (we-)intend x). In a pre-socially normative sense, I ought to do my part in what we intend, and any obligation or entitlement that might come to play a part in shared intentional activity ultimately arises from this pre-social normativity. [21]

Thus, the fact that a collective agent has a membership policy that allows for abandoning the group any time a members wants to only entails that the group gives each member's reasons as much weight as each member does; it reacts as Merciful Merle, but this does

21 Schmid (2009: 53)

not mean that the member does not have to communicate, even tacitly, that she is abandoning the group.

Complex groups are more likely to have an explicit policy of membership inclusion and exclusion. If, at any time, I want to quit my job as a postdoctoral fellow at the University of Ruritania, I only have to communicate my will to the right members (someone from the personnel department, usually), and assess the costs of leaving the University. For example, I might be obliged to communicate my intention to leave one month before my departure date; but even in this case, I can freely leave, paying the costs (a fine of some kind, for example).

Thus, membership consists in a social commitment between (at least) two individual agents) or, if the group has already been created, between an individual and the collective agent. Moreover, small groups can also be members of larger groups; in this case, the smaller group would act as an agent, acquiring a social commitment towards another. This is the topic of the next Section.

6.2. Practical commitments of collective agents

My aim, as I stated above, is not to provide a complete description of what a collective agent, of any kind, amounts to. Instead, I assume that collective agents are capable of having collective goals, collective reasons, and collective beliefs about what they ought to do. Furthermore, they have collective intentions or, at least, they act *as if* they had intentional states—which, from a functionalist perspective, amounts to holding such states. I am aware that these assumptions are controversial, for the explanation of how collective agents come to hold mental states such as beliefs, goals and intentions is a source of dissent in the philosophical debate. I thus endorse Tuomela's claim that joint intentions have normative and volitional force, just as individual intentions:

> Joint intentions, in contrast to joint wants and desires, can be regarded as joint commitments to action, viz. the participants' interdependent commitments to perform their parts of the joint action and their responsibility for the total joint action getting performed. (We can also

> say more generally that the participants are jointly committed to reach a joint goal—such as X's having been performed jointly by them—and their plan to achieve this joint goal, as reflected in their part-performances.) Such joint commitments are appropriately persistent and, especially, are not consummated before the agents have jointly achieved what they we-intend (or achieved consensus about the unachievability of the intended goal). (Tuomela, 1996; 495).

Individuals can build up many kinds of collective entity by becoming its members. Many collective agents, however, are made up when the collective action is required, and they do not continue existing after the action is done. For example, suppose that Ann and Bob are walking in the park, and they arrive to a basketball court. There is people playing there, and Bob and Ann decide to join each a different team and play against each other. Bob and Ann belong to different groups, each of them having a goal that is shared by its members—specifically, the goal to win the match. This is something an individual alone cannot do; it is a collective goal, and requires collective action to be achieved. However, the two collective agents (Ann's team and Bob's team) do not continue to exist after all their members are gone, at least as the same collective agent. Thus, Ann and Bob's membership is a social commitment with the other players to promote the teams goal: to score as many points as possible, and to try to win the match. But I do not believe that each team can easily adopt other practical commitments. Maybe players can choose to follow a specific strategy, and this is a practical commitment of the collective agent nonetheless—namely, a commitment to perform the means to achieve their end. But I do not believe they can create commitments outside the barriers of that particular basketball match. They cannot adopt other goals, nor become socially committed to other agents. It is a spontaneously created team, and ceasing to be a member is a simple as walking away while saying “Great game, I have to leave now. Bye!”.

More complex and stable collective agents, on the other hand, are able to assume new goals, to acquire practical commitments, and to become socially committed to other agents. If a collective agent is not able to keep a record of its collectively accepted normative judgements, beliefs and practical reasons, then normative requirements of rationality cannot be applied to it, for there are no elements to be normatively linked.

Group deliberation is subject to the same constraints that individual deliberation: given the reasons one has, one endorses a normative practical judgement about what ought to be done. Rationality requires that intentions and actions do not contravene one's normative judgements (principle of enkrasia); it requires as well that one does not hold inconsistent intentions (principle of resolve). Those normative requirements also apply to collective agents. Furthermore, a collective agent should also judge following the directions of its past judgements[22]: it should be internally consistent. The following Figures show the enkratic and the resolve requirements applied to collective agents.

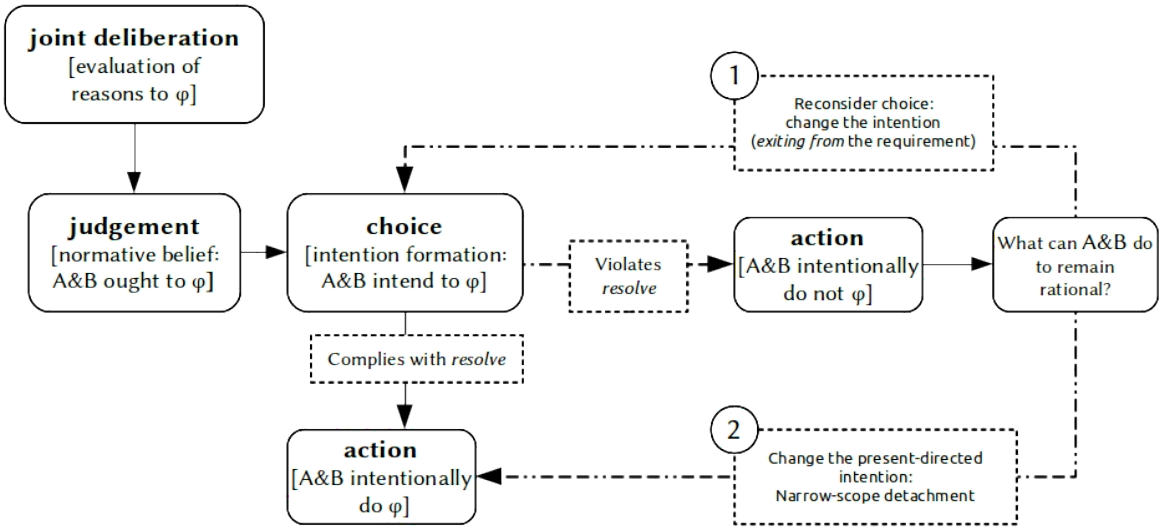

Figure 5: The enkratic requirement applied to a collective agent A&B

Here, A&B are the collective agent whose members are A and B. If they jointly adopt a collective goal, then this goal is attributable to them as a plural subject[23]. A collective goal does not necessarily require that neither of the members is individually able to attain it. For example, my flatmates are jointly committed to water the house plants—but watering a plant is something any of them can individually do. However, the obligation to water the plants falls on them as a group, because they have jointly decided to take care of the plants collectively. Similarly, a self-employed person can do pretty much everything a

22 Pettit provides an extensive argumentation about why groups ought to be constrained by their past judgements (2003b); (2007a); List and Pettit (2011).

23 See Gilbert (2006); see also Westlund (2009).

company can do, but she has to do it individually. Thus, collective goals do not necessarily require collective action, understood as an action that can only be performed by a collective agent, such as playing a football match. A social commitment to perform the group's goals (membership) normally entails a social commitment to do one's part in the plan to achieve the collective goal.

Therefore, if the group A&B have decided that they ought to do φ, the are rationally required by enkrasia not to contravene this judgement: acting contrary to the group's normative judgements would be a form of collective enkrasia. Once that A&B state that they are being akratic, they can either change their intentions, and start doing φ, which would not violate enkrasia (option 2); or they can reconsider whether they ought to φ, and exit from the enkratic requirement.

Collective normative judgements are normative judgements individually accepted by the members of the group in their capacity as members, and individually attributed to the collective agent. It is not required that each member believes that the group ought to do φ—only that they accept that the group ought to do φ. Suppose that David, a member of the company that has adopted the marketing strategy A, does not really believe that the company ought to adopt strategy A. However, *qua* member, David has a social meta-commitment to promote the group's goals, and thus he accepts that the company ought to do φ. Acceptance, against belief, can be intentionally done[24]: David believes that it is correct to assert that the company ought to adopt strategy A, because he accepts the proposition expressing the normative judgement[25]. If David did not accept the normative judgement, he probably would have to cease being a member.

On the other hand, collective intentions are subject to the resolve requirement:

24 See Tuomela (2000); Bouvier (2004).

25 Tuomela's concept of we-attitude gathers this notion of acceptance:

> [C]ollective acceptance amounts to coming to hold and holding the right kind of achievement-oriented collective attitudes or shared we-attitudes involving a substantial amount of collective commitment towards the sentences (or propositions) accepted for the group. . Tuomela (2002: 151–2)

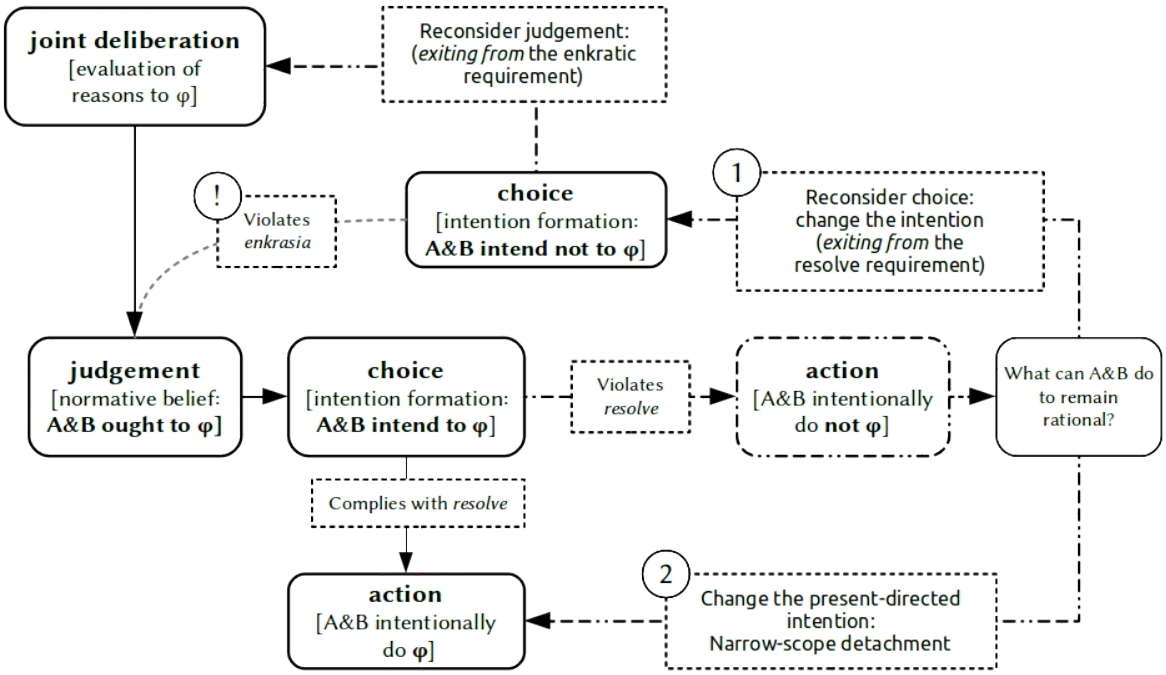

Figure 6: The resolve requirement applied to collective agents

In order to avoid holding inconsistent intentions, a collective agent can revise one (or more) of its previous intentions in order to change it and avoid irrationality. However, it is important that the change of intentions does not violate enkrasia.

Hence, collective agents are capable of acquiring practical commitments. There are two basic ways of acquiring this kind of commitment: either through tacit attribution, or through explicit deliberation by its members. In our example above, merely by joining a basketball team in order to play an informal match is a way of implicitly accepting to promote the team's goal: to win the match. In many cases, neither membership nor collective action is preceded by deliberation. This does not entail that the group does not have a goal—it has one, and it is known by its members, who acknowledge that participating in the collective activity is a reason for them to perform their part.

Second, a practical commitment to achieve a goal can be the result of a deliberation process in which the members are involved[26]. The process of collective deliberation is very similar to that of individual deliberation: assessing the options, and evaluating the reasons the collective has for achieving one (or more) of the available goals. Once there is an agreement about what the collective agent ("we") ought to do, the members acquire, in virtue of their general meta-commitment to do their part in the achievement of the collective goals, a social commitment to do their specific task.

Collective deliberation, and the process through which decisions are made, can be very simple and easily obtained from individual reasons, commitments and reasons. For example, a basketball team can reach a quick consensus about what strategy to follow next[27]. But collective deliberation can also be an opaque and complex process. Suppose that a big company is deciding what marketing strategy to follow. There is a variety of options, and there is a great amount of dissent amongst the members. Depending on the decision procedure that the company has adopted, the result can vary; and some procedures reflect in a more suitable way the group's previous commitments. This is the central claim behind the *discursive dilemma*, put forward by List and Pettit[28]. The discursive dilemma shows that there is a problem of logical consistency among individual judgements, and the collectively accepted judgement that results from the sum of individual ones. Judgement aggregation does not always entail that there is a coherent relation between the members' judgements and the collectively accepted conclusion. It may happen that most members agree on most of the premises, but they do not individually endorse the conclusion; they have to decide whether they are going to decide given the acceptance of the premises, or the acceptance of the conclusion. List and Pettit[29] show that no procedure can warrant four desirable properties:

26 Although not necessarily through a consensus amongst every member; operative members, i.e. those whose role allows them to do so, have the capacity to decide collective goals.

27 Although, in this case, it can be argued that the coach is the operative member, and the players are only following orders; but let's imagine a team in which every player is involved in the decision-making process.

28 Pettit (2001); (2003b); List and Pettit (2006); see for an overview List and Puppe (2009).

29 List and Pettit (2002); (2004).

> [T]here exists no judgment aggregation rule satisfying four conditions: universal domain (all combinations of rational individual judgments are admissible as inputs), collective rationality (only rational collective judgments are admissible as outputs), anonymity (the aggregation is invariant under permutations of the individuals), and systematicity (the collective judgment on each proposition is the same function of individual judgments on that proposition). [30]

Thus, there is no decision-making procedure that is able to warrant that the group will not make inconsistent judgements. Inconsistency between the sum of individual judgements and the resulting collective judgements cannot be avoided through any voting procedure.

The discursive dilemma shows that collective decision process take two different things into account. On the one hand, the inputs of the decision process would be the reasons each member has, that serves as a premise that counts for, or against, certain conclusion. On the other hand, the procedure itself is central to the relation between the premises and the conclusion, and thus affects the conclusion as well. Suppose that 49 out of 50 members of the company committee support one of the available options, the marketing strategy A. If the committee has previously decided that the decision procedure requires to make choices only by unanimity, then the company cannot conclude that it ought to follow the marketing strategy A.

This is the main difference between individual and collective decision processes. Individually, one usually judges that one ought to do what one has most reasons to do. Of course, there are still important similarities, because it is not always easy to compare facts in order to know what reasons are better, or stronger. An individual agent can have doubts about whether to choose option A or option B. If this is a crucial choice, then she will not choose unless the balance of reasons clearly points towards one of the options. She can decide a decision procedure: toss a coin, or choose alphabetically. The decision procedure serves as an exclusionary reason to choose: it affects other reasons, although it does not directly provide a reason for doing what has been chosen. The same goes with collective agents. It can be said that a company has decided to implement the marketing strategy A instead of B because every member in the committee voted for it; or that the company has

30 List and Puppe (2009: 458).

decided to implement the marketing strategy A instead of B because of reasons α, β and γ. Both the member's reasons and the decision procedure serve as collective reasons.

Lastly, insofar collective agents are capable of acquiring practical commitments, they are can also become socially committed to other agents, both individual and collective. When they adopt a social commitment, collective agents accept that commitment as a reason to perform the content of the commitment: in this sense, they function as individual agents.

Concerning the relation between collective agents and collective actions, a group can commit itself to another agent to do something (thus engaging in a *social* commitment) that does not require that the entire group, not even a subgroup, performs the action: "operative" members can sometimes fulfil the commitment[31]. For example, my sister and I can commit ourselves, as a group, to paint a friend's house. Painting a house does not require collective action: I can paint a house on my own. Thus, either if I paint my friend's house, or my sister does it, or we do it together, we would still be fulfilling our commitment. However, other social commitments do require that all members are engaged in the performance of a collective action. Suppose that a fire brigade formed by five firemen are are required to extinguish a fire. The situation is such that all of them are required to participate in the action; otherwise, the fire could not be extinguished. In this case, the individual actions that make the commitment to be fulfilled or violated differ from those entailed in the example of the house being painted.

Although the fulfilment condition may be different for a social commitment whose debtor is a collective agent, and that whose debtor is an individual, the remaining normative elements remain analogous, as well as their structure. Suppose that the collective agent A&B, which is a group of boys who live in the same neighbourhood, has accepted an request from Elisabeth to prune the trees in her garden[32]. Having accepted the assignment is an exclusionary reason for A&B to prune the trees. Furthermore, by

31 Tuomela (2002).

32 For the sake of simplicity, I will assume here that it is not a case of two interdependent commitments: Elisabeth does not have to do anything in exchange.

accepting the assignment, A&B let Elisabeth know that they accept the social commitment as a reason to prune the trees, a reason over which Elisabeth has acquired certain normative powers. For example, she can release A&B from its duties. A social commitment adopted by a collective agent does not necessarily require collective action. After all, either A or B can prune the trees, and the commitment would still be fulfilled. However, if the commitment is cancelled by the debtor, the agent (A&B) can be held responsible for the outcome (the trees not being pruned), and also its members (A and B).

In sum, an agent is able to acquire practical commitments if the following conditions apply:

(i) It is able to acknowledge and accept practical normative judgements.

(ii) It is able to guide its behaviour through those judgements.

(iii) It is able to recognize inconsistency amongst its judgements, choices and actions, this is, is able to respond to the requirements of rationality.

Collective agents are capable of (i)-(iii): they aim at collective goals which they intend to achieve, and they are capable of assessing whether their actions contravene their goals and intentions. Without (iii), which amounts to a feedback system, collective agents would not be able to know whether the collective goal has been attained, either. And it is because of these three conditions that collective agents, such as A&B, can be appropriately be held responsible. This is the topic of the next Chapter.

CHAPTER 7. COLLECTIVE RESPONSIBILITY

It is quite frequent, in or everyday speech, to attribute responsibility to collective agents. Banks, or even the financial system as a hole, bear (at least some) responsibility for the present economic situation; the BP corporation is responsible for the *Deepwater Horizon* oil spill in the Gulf of Mexico; a Red Cross team is responsible for saving a man's life. Insofar groups of individuals are able of engaging in collective action, the consequences of their actions, as well as the actions themselves, can be collectively attributed to them. I will not address here the problem of whether collectives can be held *morally* responsible[1]. My aim is to highlight the role of collective commitments in the explanation of an outcome, but I leave open the question of whether collective moral responsibility can be founded in responsibility as attributability.

The structure of this Chapter is as follows. In §7.1, two problems of responsibility for omissions will be discussed. First, it is not clear whether any collection of individuals can bear collective responsibility when they fail to act jointly. Second, even if responsibility is being attributed to a collective agent, the relation between that agent and the outcome is not clear, as I argued in §5.3.2. In §7.2, I will discuss the problem of the relation between collective and individual responsibility, this is, under which conditions and criteria collective responsibility can be individually distributed amongst the group members.

1 I believe they can; insofar collectives are suitable for being held responsible in the sense of attributability, it seems natural to think that they are also suitable for being morally accountable, if their actions have a moral (either positive or negative) status. However, regarding associated social practices such as loss compensation or punishment, it can be argued that collective agents cannot be punished in the same sense that individuals can, because they do not respond to reactive attitudes, and thus the role of punishment in regulating behaviour would fail.

7.1. Attributing responsibility to collective agents

In Chapter 5, I argued that attributions of retrospective responsibility for an outcome require the outcome to be explained in terms of the agent's actions and agential capacities. This does not necessarily entail that the agent has consciously and voluntarily performed the action, but it does entail that the agent stands in a particular normative relation with the outcome. Explanations always take into account what is normatively and empirically expected, and thus they are required and offered against a normative standard that describes how things should work, and how people *ought* to behave. The obligation entailed by that *ought* is not necessarily moral, but *rational*: what people ought to do given the reasons we can justifiably attribute to them. Stating that "agent A is responsible for outcome O" is analogous to asserting that "outcome O is the case *because of* agent A". Therefore, responsibility attributions involve some kind of causal or explanatory judgement.

Attributing collective responsibility, then, consists in appealing to a collective agent in order to explain the outcome for which that agent bears responsibility. Omissions are specially problematic for responsibility, and the role of the agent's commitments acquire greater relevance in the explanation of the outcome. I will argue now that it is misleading to attribute collective responsibility to any collection of individuals who are able of collective action, and I will suggest an approach to collective responsibility that takes into account the collective agent's social commitments to other agents.

7.1.1. Bystanders and members

It has been argued that collective responsibility, specially that derived from omissions, can be attributed to any collection of individuals who are, in principle, able to create a group in order to act collectively. Consider the following example:

> **Coordinated Bystander Case**: Four bystanders are relaxing on a riverbank when six children on a raft run into trouble when they and their raft end up in rapids. They are hurtling helplessly toward a dangerous waterfall downriver and are unlikely to survive if they go over it. Nothing any of the four bystanders can do as an individual will make a difference, but there is an obvious course of coordinated action they could take to divert the raft into calmer waters. This measure would pose little risk to the bystanders and would save all of the children. [2]

Are these four bystanders collectively responsible for failing to form a group to save the children? Isaacs argues that they are. Her strategy consists in showing that the collection of bystanders form a *putative* group[3], which is a group whose members belong to the group by aggregation, but that is able to become a group by affiliation, given certain conditions. In the example, above, the four bystanders belong to "the group of people that is witnessing the raft events", but they could engage in collective action, which would turn them into members by affiliation. Let's suppose that every individual knows both what collective action ought to be performed, and that there is common knowledge both about the solution and of the other agents' awareness. May argues that if those two conditions are met, the group qualifies for being the target of collective responsibility, because it incurs either in a collective inaction or a collective omission:

> If people are able to decide how to act as a group, and they decide not to act, then the failures to act constitute collective omissions. If people are able to decide how to act as a group, but they do not reach any decisions, and as a result nothing is done, then this is a clear case of collective inaction. But if the collection of people has no previous history of collective action or if it is not clear whether they could have reached a decision at the present, then it is not clear that the lack of action by these people is an instance of collective inaction, the consequences of which the group can be responsible for. Collective inaction does not merely involve aggregated individual inaction, with the mere possibility that collective action could have occurred. For inaction to be collective there must be some sense in which the group as such failed to act. There must be practical plausibility to the counterfactual claim that the group could have done otherwise. [4]

Other authors go as far as claiming that, in principle, every group by aggregation can bear collective responsibility. For example, Silver claims that present-day Americans can share responsibility for actions associated with American slavery despite the fact that no one

2 Isaacs (2011: 143)

3 The concept of putative group is drawn from May (1992: chap. 6).

4 Ibid., 108.

alive today participated in those actions[5]. Similarly, Held argues that the inhabitants of a country can be held responsible for omitting to overthrow their political system, because they fail to constitute themselves as a group through adopting a decision procedure in order to act collectively[6].

This argument can also be applied to outcomes resulting from large-scale actions, such as wars, world poverty, or global warming. However, I do not agree with the claim that, merely by *having the possibility* of engaging in collective action, a collection of individuals qualifies for collective responsibility. This is not to say that those individuals are not responsible at all, or that they are exempt. They can share responsibility for the outcome, but they do not constitute a collective agent, which is the proper subject of collective responsibility. The kind of membership required to belong to the collective agent provides a better criteria in order to determine whether an outcome is attributable to a group as an agent, or as a collection of individuals[7].

First, the distinction stated in Chapter 6 between *collective obligations* (which require membership as affiliation) and *general restricted obligations* (which only needs an aggregation of individuals) is fundamental to understand collective responsibility. Some analysis of collective responsibility use examples that are in fact general restricted obligations, specially when it comes to analysing how to distribute responsibility among individuals belonging to spontaneously created groups, such as a collection of bystanders to which a general obligation is attributed, for example, the obligation of intervening when someone needs help, which includes the obligation to create a group if needed, and to jointly act as a group.

Second, it is important to distinguish between collective and shared individual responsibility. The former refers to responsibility attributed to a collective agent. Shared individual responsibility, on the other hand, is not necessarily derived from collective responsibility. It also refers to the individual responsibility for the production of an

5 Silver (2002).

6 Held (1970).; see also Tännsjö (2007).

7 See Petersson (2008).

outcome, which results of the sum of the individual contributions—global warming, world hunger, discrimination against women, and so on. In these cases, the causal contribution of the agent is evaluated under the light of her role in the collective, or in the production of the outcome. For example, we as individuals share the responsibility of discrimination against women, given our actions that promote this outcome, and our inactions to prevent a case of discrimination (given what we can do).This is not to say that discrimination against women is merely the result of many individual contributions. Certain laws support this kind of discrimination, and those laws are introduced, passed and enforced by collective agents[8]. Thus, the explanation of why women suffer from discrimination includes both individual behaviour and institutional facts caused by collective agents, because both individuals and institutions are normatively expected[9] to value non-discrimination, behave non-discriminatingly, and to promote non-discriminatory policies. Shared responsibility, on the other hand, can also be derived from collective responsibility, as I will argue in §7.2.

Hence, the difference between membership as affiliation and membership as aggregation is central to the problem of attributions of collective responsibility, and to the distribution of individual responsibility amongst the group's members. Following Mellema, collective responsibility requires that the individuals who belong to the group are members by affiliation:

> [C]ollective moral responsibility is not something that affects innocent bystanders in the manner of a contagious disease. Rather, a person must do something or omit doing something to qualify for membership in the collective. [10]

Therefore, regarding omissions, *individual failures to act jointly* should be distinguished from *collective failures to act*. The kind of responsibility involved in each case differs. I will

8 The boundaries of the collective agent that produces a law may be fuzzy, or can vary depending on how broad or narrow we construe the concept of collective agent. It can be argued that the Parliament is the collective agent in question; but the citizens who have voted the ruling party, specially those who did so because they expected these discriminatory laws to be introduced, might be good candidates to members by affiliation, and thus belong to the collective agent.

9 Not *universally* expected, but some people, such as me, expect it.

10 Mellema (2003: 129).

turn back to this point below, when discussing the distribution of collective responsibility amongst individual members.

To sum up, the concept of membership as affiliation allows to mark the limits of responsibility, specially concerning omissions. While bystanders may be individually responsible, only members through affiliation can be collectively responsible.

7.1.2. Collective responsibility: the case of collective omissions

Collective responsibility requires that the outcome which is the object of responsibility to be explained, at least partially, appealing to certain agential features of the collective. This claim of course entails that the collective must possess those features, which are the same applying to individuals in order to be appropriate targets of responsibility attributions. I explained the agential requirements of responsible agents in §5.2.1. In summary, the agent must fulfil both freedom conditions, which allows the agent's capacity of control over her own actions, and epistemic conditions, in particular, reason-responsiveness, and the actual display of self-control. Thus, attributions of responsibility require:

(i) that the agent meets certain control and reason-responsiveness conditions;

(ii) that the outcome is suitably explained in term's of the agent's decisions and actions (or omissions), so what is expected that the agent will do or decide contrasts with what she actually did or decided;

(iii) that the agent can be held responsible for her own actions and decisions, so when tracing prior conditions, the agent can be held responsible for those conditions as well.

I argued that for (iii) to be attained, a normative account of agency is needed, such as the legal standard of reasonable agent. If the agent has acquired previous commitments, these

commitments (if known by the agent who demands an explanation) also serve as a normative standard, because they derive from the normative reasons the agent has[11].

In Chapter 6 I have shown that collective agents are capable of acquiring practical (internal) commitments to attain a collective goal, and that because of this capacity they can also acquire social commitments to other agents. Condition (i) above, which refers to the agent's capabilities, has been expressed by Pettit through the following two criteria:

- Value judgment.—The agent has the understanding and access to evidence required for being able to make judgments about the relative value of such options.
- Value sensitivity.—The person has the control necessary for being able to choose between options on the basis of judgments about their value. [12]

Putative groups mentioned above do not meet these requirements, and therefore they do not meet the necessary conditions to be held collectively responsible. However, despite the kind of group which is the target of the responsibility attribution, omissions still pose a problem. If nobody did anything, why does the kind of group matter? As Larry May pointed out, the analysis of collective responsibility for omissions inherit two different problems:

> Why, among the countless things members of a group fail to do, should certain failures be singled out as constituting "collective inactions"? And how should blame for the harmful consequences of these inactions be apportioned within the group? [13]

First, as I argued in §5.2.2, omissions are problematic because the relation between the agent and the outcome is not evident: there has to be a link of some kind that connects agents with outcomes; otherwise, every non-action would constitute an omission, and everyone would be responsible for everything (see §5.2.2). On the other hand, the problem of how to distribute responsibility among members, or whether responsibility voids are possible, are central issues in the debates on collective responsibility. There is a third problem halfway between these two questions: how can the limits of a group be identified, when no action has taken place? This is, if nobody does anything, how can

11 Of course, it is also possible to evaluate the agent's commitments in the light of other normative standards. But my aim here is to argue that the agent's commitments serve as those normative standards, so I will leave this possibility aside.

12 Pettit (2007b: 175).

13 May (1990: 270)

members from non-members be distinguished in the attribution of collective responsibility?

Taking prospective collective responsibility for an outcome means that the group acquires a commitment to perform the necessary actions for its promotion. However, this commitment can be fulfilled in many ways. It may necessitate the action of every member or just some of them. Furthermore, it is not required that the set of actions needed to achieve the outcome form themselves a collective action: a group can take responsibility over an outcome that requires the action or the omission of only one of its members.

To illustrate this, let's bring back the problem of Queen of England (§5.3.2). The Queen of England problem, proposed by Sartorio[14], shows that it is problematic to hold a promise-maker, rather than anyone else (such as the Queen of England), responsible for the consequences derived from failing to fulfil a promise. The promiser has failed either to fulfil the promise, or to perform the necessary steps, which in this case, amounts to watering some plants in order to keep them alive.

Suppose that my flatmates Ann, Bob and Carol promise me to water my plants while I am on vacation. I go on vacation, and when I come back, my plants are dead. Had any of my flatmates watered them, my plants would still be alive. Then, who is responsible for my plant's death? Either my flatmates are responsible as a group, or every flatmate is responsible individually for her omission, due to the fact that all of them failed to water my plants. I will argue that the failure of any specific flatmate to water the plants do not play a role in the process explanation of my plants' death, but their failure as a group does.

The group in exemplified above is very loosely structured. It differs from a collection of bystanders in that they have collectively agreed to water my plants, while the responsibility to help others, for instance, is a restricted general obligation, and does not need for an explicit agreement. In the case of more structured groups, the distinction between the individual level (responsibility as member) and the group level (collective responsibility) is more evident: in my example, it is possible to argue that the group as

14 Sartorio (2004).

such has been formed in the same moment as the promise is made[15]. My argument would remain the same if applied to more structured groups, such as gardening companies. But, in the case of having hired a gardening company, there would a contract, and therefore a legal system, that would affect the normative structure of the situation. Although the practice of promising is also mediated by social norms, I think it simplifies the argument. Therefore, I will take the risk of sticking to this simpler situation.

There is not any link between the death of the plants and the actions of any particular flatmate, but between the state of the plants and their commitment as a group. Imagine for example that Ann waters the plants. When I come back, my plants are alive. In this scenario, neither Bob nor Carol has watered them, but the group's social commitment is fulfilled. My plants are alive because my flatmates have watered them (as they promised)[16]. Now, in a different scenario, none of them waters my plants, and they die. Bob and Carol, however, stand in the same causal relation with the plants than in the previous scenario. Thus, their causal contributions do not make a difference in the attribution of responsibility. In the first scenario, Ann's causal effectiveness makes a difference, and this difference has effects in the attribution of responsibility to the group; so it seems plausible to assume that Bob and Carol stand in a particular normative relation to the group (their status as members) that makes appropriate to claim that they have fulfilled their commitment, although they have not watered the plants themselves. This also excludes other agents that do not stand in this particular normative relation, such as Queen of England, or any other innocent bystander.

The opposite situation is also possible. In the first scenario, Ann's action leads to the fulfilment of their commitment; but let's imagine that all three water the plants, and when I come back, the plants are dead because the excess of water. They would have also failed, as a group, to take care of the watering of my plants. Individual actions can be

15 Engaging in a collective promise would qualify this group as an incorporated group, differentiating it from collections of individuals see Stilz (2011).

16 Of course, I could say that my plants are alive because Ann watered them, but Bob and Carol have not violated any commitment; and I suppose that, were Ann not able to water them, either Bob, Carol or both of them would have taken care of my plants.

coordinated in many different ways to achieve the goals of the group[17]. Their failure to achieve the goal does not merely depend on the causal contributions of each member, even if the action required can be performed individually, such as watering the plants.

My flatmates have taken prospective responsibility over my plants as a group: no specific member of the group has committed her will to watering the plants, although they all have accepted being members of the "watering group". However, their membership entails a generic meta-commitment to perform the group's goals; and, in this case, my flatmates were fully aware of what this goal was, and through becoming a member, they have also acquired a social commitment amongst them to water the plants.

As I argued in §5.3.1, I expect that my flatmates explain and justify to me why my plants are dead. They all stand in a particular relation with the plants (in virtue of their promise) that normatively requires that the plants are watered, and this makes their omission as a group as informationally relevant in a causal explanation of my plant's death, but not every particular inaction is relevant to the explanation. As I have argued, some inactions count as omissions (and not merely one of the uncountable non-actions of an agent) only in virtue of their relation to the group's prospective responsibility derived from the collective commitments acquired.

As a last remark for this section, I will briefly return to the problem of distinguishing members from non-members in the case of collective omissions. Causal accounts of responsibility for positive actions (this is, those in which the agent actually does something which fulfils or violates her commitments) delimit the group by assessing the individual contribution to the outcome. This narrow account has some difficulties in explaining the responsibility of members who belong to the group but have not causally contributed to the outcome, on the one hand, and in evaluating the individual responsibility for a group omission. A solution to this puzzle can be found in the concept of membership stated above. As long as it presupposes a kind of meta-commitment to perform the appropriated actions to fulfil the group's prospective responsibility, the

17 This coordination can be analysed as a kind of weak collective agency; see Petersson (2008).

boundaries are to be found at the individual level, in the social commitments entailed by membership. Following our example above, checking whether the plants have already been watered is an individual meta-commitment: it does not stand in a direct relation with the state of affairs, but it does stand in a particular relation with the goal of the group. Furthermore, it is possible that not every individual stands in the same membership relation to the group, and so the distribution of responsibility has to take into account this variation in the degrees of implication of each member.

7.2. Membership: the distribution of responsibility

Belonging to a collective agent carries certain prospective responsibilities, acquired through the social commitment in which individual agents incur for becoming members. The members' meta-commitments to the group's goals are the link between collective and individual retrospective responsibility. However, the transition from collective to individual responsibility is problematic in several aspects.

7.2.1. The dilution of responsibility

First, even if collective entities act through their individual members, collective actions do not necessarily coincide with their member's actions. In our example above concerning watering my plants, these actions coincide: the group of my flatmates has watered my plants, and it may be the case that each flatmate has watered the plants. But let's suppose that the collective agent is a basketball team. Although the team has win the match, none of its members has individually won the match herself. This is an action that can only be attributed to the team as a whole, which has been performed through each member playing her part. What is the relation between the collective responsibility of winning a match, and the individual responsibilities of each member? After all, the outcome (winning the match) cannot be attributed to any of them individually. This first problem thus refer to how to relate collective actions to individual contributions.

There are two confronting views regarding the solution to problem. On the one hand, it can be argued that individual responsibility is the result of dividing and distributing collective responsibility. This view is often called the *pie-model* of responsibility[18]. From this perspective, collective responsibility can be metaphorically seen as a pie that is divided, and each individual member gets a share. Collective responsibility is *individually shared* amongst the members of the group. In the case of a basketball team, its players share responsibility over the outcome, which is the result of the match. The problem here is that, if the number of members is large, then each member receives a very small amount of pie: collective responsibility dilutes, and the larger the group, the more diluted individual responsibility is. Suppose that our basketball team has a thousand players: the impact of each player's actions to the overall result is small, and so is her share of responsibility.

On the other hand, it has been argued that collective responsibility does not dilute, but remains the same and is applied to each member. Zimmerman defends an anti-dilutionist account, and offers the following example:

> Imagine a group of ten teenagers pushing a large boulder off a plateau, so that it rolls down a slope and wrecks a car at the bottom. Each of the teenagers intends to contribute to the damage to the car and freely participates in the enterprise, in the full knowledge that his contribution to the enterprise is required if the boulder is to be shifted and the car wrecked at all. [...] [T]his is a case of standard simultaneous group action. Now, who is morally responsible for the damage to the car and to what extent? I have no doubt that many would say, given the facts of the case, that the ten teenagers share the responsibility for this outcome of their group action. But this seems to me false, if the suggestion is that none of the teenagers is fully responsible, that is, that each of them has an excuse such that his responsibility is diminished. I believe, on the contrary, that each of the teenagers is fully morally responsible for the damage to the car. [19]

Zimmerman claims that there may be different excusing conditions for each teenager. However, this would not affect his argument: that no member is automatically excused, and therefore her responsibility is diminished, merely because the outcome attributed is the product of collective action. Furthermore, Zimmerman argues, contributing to the

18 Mellema (1985); see also Lenk and Maring (1991); Tollefsen (2006).

19 Zimmerman (1985: 116).

collective outcome makes any member responsible, even in cases in which each member's actions were neither necessary nor sufficient. This is, that the outcome would have be brought about regardless of any particular member's contribution—although a minimum amount of contributions is necessary (and sufficient). The main idea is that collective responsibility can be directly transferred to individual members.

Anti-dilutionist accounts face the problem of attributing full responsibility in large groups. Suppose that we apply Zimmerman's account to the *Deepwater Horizon* oil spill: is every member of BP fully responsible for the spill? Zimmerman would argue that some individual agents could be excused (for example, because they work at the lower levels of the organization and there is nothing they could have done to prevent to spill); but, as I argued in Chapter 5, excuses affect accountability, although accepts attributability. It could be possible that certain members, given their role within the organization, are in fact exempt. However, anti-dilutionism does not address the question of how the outcome is explained in terms of the collective agent, or in terms of its members[20]. It seems false to assert that the oil spill happened because of member A's actions, as well as because of member B, member C and so on. It seems more plausible that members A, B, C, and others, *share* responsibility for the oil spill.

There is a third view on the distribution of responsibility, consisting in denying that responsibility is something that can be shared. Even if it is accepted that collectives can be held responsible for the outcome of a collective action, individuals can only be held responsible for their individual contributions. I agree in that individual members are responsible for their contributions; but it is problematic to open a wedge between collective and individual responsibility. Bringing back our example above, this view would claim that BP is (collectively) responsible for the oil spill, but none of their members is responsible for the oil spill: each member is responsible only for their own actions, none of

20 Things would be different if moral responsibility did not require attributability. Although I believe it does, it is possible to argue for an anti-dilutionist account of blameworthiness. The causal relevancy of each factor may be difficult to determine; however, blame, and specially punishment, are quantifiable. Thus, it might be the case that each member is responsible (as attributability) at different levels, but that all members are equally accountable.

which is "spilling oil in the Gulf of Mexico". This account is misleading, because of two reasons. First, the outcome resulting from a collective action seems to float over the members' heads, without touching any of them. This is: far from constituting a few exceptions, responsibility voids would be the rule. Second, it makes certain members automatically exempt. Suppose that a certain member of the teenagers conspiracy group above, called Fred, has a task: to keep watch in case the owner of the car arrives, and to warn the others so they are not caught. As a matter of fact, the owner does not arrive. Fred does not even touch the car. If we aim to defend that collective responsibility is not shared, then Fred is not at all responsible, for he did nothing to causally affect the car. And this conclusion is not one that I would want to endorse; thus, there is in fact a relation between collective and individual responsibility.

In spite of seeming attractive, dilutionist accounts are problematic because they leave the door open to the possibility that individuals share a little amount of responsibility, if the outcome has been produced at large-scale. Furthermore, it also has to deal with the problem of considering responsibility as a kind of "substance":

> But it is a mistake to think of responsibility as a substance of which there is a fixed amount, such that, if the collective takes its share, then there is less for the individuals. The fact that a collective is responsible for some action does not mean that the members are not responsible for their contributory actions as well. While the answer to the question "Why did you (all) do x?" will refer to the collective perspective, we can still ask an individual agent "Why did you contribute to doing x?". Her answer to this question must be in terms of her own perspective. When a member acts so as to contribute to carrying out an action of the collective, she is (or ought to be) fully aware of what she is doing. And, as an individual moral agent, she is free to reject and oppose the actions of the collective of which she is a member. If she chooses not to do this, she is fully responsible for her contribution to the collective action. [21]

I believe that members *share* individual responsibility if the collective to which they belong is responsible, in the same sense that they share a causal role in the explanation of the production of the outcome. They contribute to the collective goal's achievement. Becoming a member, I have argued, is a social commitment between an individual and a group (or between two individual agents if the group if being created at that moment).

21 Mathiesen (2006: 250)

Affiliation requires that the agent accepts that being a member is a normative reason to promote the group's goal or goals. In this sense, every member accepts the fact that they have acquired a practical commitment as a reason to perform the content of the commitment (see §6.2). If a member does not accept the collective goal, then she is a kind of outsider: in order not to violate her social commitment as member, she ought abandon the group. However, as long as she is member, she shares responsibility for the group's actions. Pettit puts it in the following way: "the members will have responsibility as enactors of the corporate deed, so far as they could have refused to play that part and didn't"[22]. Thus, membership requires acceptance: not only the member accepts to be a member, but also accepts the collective goal and its promotion.

To sum up, the acceptance involved in being a member provides the member justificatory reasons: she promotes the collective goals *because* the collective has decided to achieve these goals, and she belongs to the collective (for whatever reasons). Let me illustrate this with an example. Mary needs money, and she cannot find a job as a philosopher (how surprising!). She looks for jobs all around the city, and she ends up by being hired by the King of Burgers, which is a fast food company that sells burgers and other high fat food. Mary would never eat in the restaurant she works for: she basically believes that she is selling a sandwich made with low quality meat and a load of fat. She would never recommend to any of her friends to have lunch there. It is quite clear that she does not personally believe that she ought to offer a dessert with each meal. In spite of all these facts about Mary, she belongs to the King of Burgers organization, and she accepts the following collective goal: "we should do our best to sell the most expensive menu". This goal guides her behaviour: every time a customer arrives, she offers the biggest menu, and the possibility of purchasing a dessert. Does Mary hold any responsibility for distributing low quality and high fat products? I would say that, even if she does not identify herself with the collective goal, and she does not personally believe that the collective goal ought to be attained, she has accepted to be a member, and she thus

22 Pettit (2007b: 192).

promotes the company's goals. If King of Burgers can be held collectively responsible for distributing unhealthy food, Mary shares responsibility, with the other members, of that same outcome.

Of course, the role of each member in the collective group will determine whether a member plays a central role in the promotion and production of the outcome, or has only access to little or none power of decision within the organization. Different roles, degrees of control and access to the deliberation mechanisms of the collective will surely affect distributions of responsibility. Analysing these differences exceeds the scope of this work; my point is that, merely by being a member by affiliation, and by promoting the group's goals, an individual shares responsibility over the resulting outcome. Therefore, as I will now discuss, responsibility voids are rare and exceptional, at best.

7.2.2. Responsibility voids

Pettit argues that the discursive dilemma shows that it is possible for a collective agent to make a collective decision whose consequences are collectively attributable to the group, whereas it is not attributable to any of its members. In Pettit's example, the members of a committee have to make a decision about the enactment of a policy. The committee is made up by co-workers who have to decide whether they agree with a pay sacrifice in order to install a safety device. The decision procedure is such that the members will only take into account the members' views on certain selected reasons, rather than the concluding normative judgement deriving from these reasons. They have to individually evaluate whether there is a real danger, whether the safety device is effective, and whether the pay sacrifice is bearable. If they assent to these three premises, then they will assent to the conclusion: the pay sacrifice ought to be made. Each issue is decided by a majority vote. The following table shows the voting results:

	Serious danger?	Effective Measure?	Bearable Loss?	Pay Sacrifice?
A	No	Yes	Yes	No
B	Yes	No	Yes	No
C	Yes	Yes	No	No
Majority	Yes	Yes	Yes	No

Table 6: Safety Measures[23]

Each member is opposed to the pay sacrifice, but for different reasons: A does not believe that there is a real danger; B believes the device is not effective; and C cannot bear the loss. The aggregation of the different reasons results in the installation of the safety device, although each member does not individually believe the device ought to be installed. Pettit claims that the personal responsibility of A, B and C for the installation of the safety measure is diminished, given that they personally opposed to that result. Only the group as a whole can be blamed[24].

I do not agree with Pettit's conclusion[25]. A, B and C have accepted the result, and they ought to promote it—by taking the necessary steps that lead to the installation of the safety device. The members have chosen a decision procedure and reached a conclusion. The committee's decision, and the actions that follow to that decision, can be explained in terms of the members actions and agential capabilities. There are no exempting conditions: they fulfil all the requirements for being held responsible. The concept of acceptance is therefore crucial. A member does not only promote, or contributes, to a collective goal: she also accepts it, along with its normative force. They may offer whatever excuse to justify their actions; but even if each member is fully excused, they are not

23 Ibid., 197

24 Copp reaches a similar conclusion, which he calls the Normative Autonomy Thesis Copp (2006); see also (2007). However, Copp focuses on moral blameworthiness; he does not claim that individuals are not attributable, but not accountable:

> There are possible cases in which (i) individuals act in official organizational roles on behalf of collectives, (ii) the choices and actions of these individuals are entirely rational and morally innocent, or at least excusable, and yet (iii) there is moral or rational fault that must be assigned somewhere, and (iv) the only plausible candidate for the assignee of such fault is the collective. Copp (2006: 216)

25 See, for a detailed criticism of the conditions and prevalence of responsibility voids, Braham and van Hees (2011). For a criticism of responsibility voids arising from the discursive dilemma, see Hindriks (2009).

exempt. Let's bring back our example above. Mary might have dozens of excuses that justify her membership to the King of Burgers, but she is not exempt. She has acquired a meta-commitment to promote the group's goals, and it is in virtue of this commitment that she shares responsibility with all the other members of the company.

An individual exemption does change the scenario. If a collective agent is exempt, then its members are exempt as members as well—but this does not rule out individual responsibility for one's own actions. And, similarly, if every member is exempt, then there is no collective responsibility. Imagine that a group of people, A B and C, are forced to commit a collective action which causes some harm. They are not a collective agent, they are an aggregation of individuals: there is no membership by affiliation; and they are exempt as individuals, because they have been forced and therefore they do not meet the agential conditions required for explaining the outcome in terms of their goals, reasons and commitments. And, if one member among the set of members is exempt, then it can be argued that she is not a part of the collective, at least as an agent. Exemptions affect attributions of responsibility by showing that the agent did not freely exercise her capabilities, and thus the outcome is not attributable to her. In order to explain the outcome, we have to move one step further and see what has caused that this agent behave the way she did.

Thus, responsibility voids amount to cases in which responsibility can be attributed at the collective level and, nonetheless, responsibility for the same outcome cannot be attributed at the individual level. Judgement aggregation is frequently used as a paradigmatic example of the inconsistencies between the collective and the individual level. These inconsistencies would cause responsibility voids. However, responsibility voids would occur when the collective is attributable, but none of its members is—and the outcome for which the collective is being held responsible is the product of the collective action. This is, the action suitably reflects the *group's* goals, reasons and intentions, but it does not reflect the *member's* goals, reasons and intentions. Then, the collective would be attributable, but its members would be exempt. I have argued that, if all members are

exempt, it is far from clear that they are able to act jointly, and thus the collective would be exempt as well.

In sum, distribution of responsibility has to take into account the normative status of members, as well as the social meta-commitments entailed in membership, which follow from the acceptance of the group's goals and commitments. If a group agent bears collective responsibility, its members share this responsibility, and they are individually responsible for their contributions as members. My account does not exhaust the variety of relations that may hold between collective and individual responsibility. As I said in Chapter 6, there is a wide scope of groups and associations that are capable of collective agency, ranging from two people walking together to transnational organizations. The different normative structures crossing and framing them allow for different concepts of collective responsibility, and different forms of distribution within the group members.

CONCLUSION

Commitment is a pervasive concept in philosophy of action. Metaphorically, a future-directed intention resembles a hook with a chain thrown to the future, which we use to drag ourselves pulling from the chain. We have a goal in mind: we stick to our goal, and then we make our way to its achievement. Practical commitments, then, are a bond linking agents and actions, like the glue that sticks the pieces together. In this work, my aim has been to show what this glue is made of, and to provide an explanation of its sticky properties. First, it would not work without the volitional capabilities of the agent, also called *self-contro*l. Without this capacity, agents would not be able to guide their behaviour in the light of their intentions and reasons: they would be subject to the motivations they have in the present, which can be predicted, but not controlled. Thus, practical commitments are made of intentional states. Second, they contain normative elements, which are reasons and normative judgements. If our future motivations change, we still require an anchor to our previous decision that justifies that we do not change our mind if we do not have a reason to do so. The agent's motivations may be not in line with her reasons. Normative reasons are the premises from which an agent concludes a normative judgement about what she ought or ought not to do, this is, about what she has most reason to do; and the relation between those judgements and the agent's intentions is governed by certain consistency and coherence requirements imposed by rationality. Akrasia and weakness of will are two failures of rationality which are the consequence of the violation of two rational requirements: *enkrasia* and *resolve*. Hence, to be committed to a goal entails a volitional and a normative dimension.

My claim throughout this dissertation is that the normative structure of social commitments to another agent and collective commitments to a goal can be both explained using the same conceptual tools as those used in individual practical commitments. Individual, social and collective commitments share a common normative structure, which connects reasons, normative judgements, intentions and actions through the requirements of rationality, and the capacity to use our normative powers.

It is widely accepted that requirements of rationality, which are normative, but not necessarily moral, govern intentional action. They help making sense of what and why ourselves and others do. If a friend of mine tells me that she intends to register in an on-line Philosophy course, and after a few weeks she has not done it, I can ask why she has changed her mind. This question is pertinent because I presuppose a link between my friend's intentions and actions. An assertoric commitment takes place when we attribute normative reasons and intentions to the speaker. We suppose that she is under a rational obligation, given her reasons and her intentions; but this obligation is not directed towards us, but self-directed. The speaker has not created new reasons for action, this is, reasons she did not have before making the assertion; this is the main difference between assertoric and action commitments.

Nevertheless, the structure of the normative requirements of rationality can also account for the obligations and rights that arise from social commitments such as promises and requests: they are the product of the exercise of the agent's normative powers. Exercising a normative power entails the capacity of changing the normative status of things, actions, and agents. For example, if I give my sister a book as a gift, I am not merely changing the book's spatial location, from my house to my sister's. I am giving my sister the rights I had over that book: since I owned it, I was entitled to use it, to give it to someone, to read it, to throw it away. Now my sister has these rights: I have exercised my normative powers in order to change the normative relation between me and the book, and between my sister and the book. This can also create new reasons for action: now I have given the book away, there is a reason that makes it the case that I ought not

take the book without permission. As Watson puts it, "the wrong incurred in breaking a promise is the same as the wrong involved in my refusing to relinquish claims to an item I have given you"[1] . Making a promise does not merely consist in asserting one's intentions. It entails a partial loss of the normative authority an agent has over her intentions and reasons for action. The only agent who is authorised to exit from the commitment, thus revoking the validity of the normative judgement "you ought to do φ because you promised", is the creditor, which is the promisee in this case. A commitment, as I have argued throughout the previous Chapters, can be violated for very good and morally praiseworthy reasons, but this does not affect the fact that the commitment *has been violated*. Breaking a promise can be wrong from a variety of moral or legal perspectives, of course. The moral or legal maxims that state that promises ought to be kept can account of the moral or the legal wrong involved in breaking a promise. Rational requirements, on the other hand, explain why a promise is considered fulfilled or unfulfilled in the first place, which is a normative fact. Thus, my argument differs from those who argue that promises are based on actual intentions: promises *prescribe*, but do not necessarily entail, that the promiser holds a practical commitment to fulfil her promise. Normatively speaking, honest and dishonest commitments are equally binding.

Concerning collective commitments, I have argued that a collective agent requires that its members do something, or omit doing something, in order to become a group member. A mere aggregation of individuals who share some feature does not constitute a collective agent. In particular, becoming a member involves getting socially committed to the group, and to other members. Members *accept* the collective goals and are socially committed to their promotion. Collective practical commitments are thus the practical commitments of a collective agent. For instance, a basketball team can be committed to win the next match. Even if one of the players has bet that the rival team will win the game (and therefore she has reasons not to believe that her team ought to win the game), she has to accept, as a member, that she ought to promote the team's goal *because* she is a

1 Watson (2009: 16).

member, and that she ought to play the best she can *because* the collective goal is to win the match. Thus, membership entails a meta-commitment to the collective agent.

The exit conditions from the social commitment of which membership consists are similar to the exit conditions from a social commitment: the debtor has to be released by the creditor. Otherwise, if the agent does not fulfil the commitment, the latter is violated. The main difference in this respect is that the member is both the debtor and part of the creditor (the collective agent). As a consequence, it is possible that a group accepts that a member leaves as soon as this member communicates the group her wish to do so. However, this does not show that collective action and membership are not necessarily normative, in the same sense that the fact that an individual agent has normative authority over her intentions and reasons, having thus the right to exit from her practical commitments any time she has a normative reason to do so, does not entail that intentional agency is not subject to rationality requirements.

The second topic I have dealt with in this dissertation is the attribution of responsibility. Commitments are a way of taking responsibility, in the prospective sense, for a future state of affairs, whose achievement becomes the agent's goal. The relation between prospective and retrospective responsibility is often overlooked. My aim has been to connect these two concepts through the analysis of the following problem: why do the agent's previously acquired (prospective) responsibilities matter in the task of making this agent (retrospectively) responsible for an outcome? Of course, acquiring a social commitment to another agent is just one amongst other ways to put oneself, or others, under an obligation. Other obligations, such as traffic laws or moral norms, are socially attributed, and not voluntarily acquired. However, an account of the relation between prospective and retrospective responsibility is necessary to explain why unfulfilling a promise places the debtor in a specific normative relation with the outcome which results from the violation of the commitment. Because of this relation, the outcome is *explainable* in terms of the agent's actions or omissions. To attribute responsibility to an agent has a thus an explanatory component. Insofar explanations in terms of an agent's

motivational structures, and not merely her body movements, require a normative standard with which to contrast the actual behaviour of the agent, practical, social and collective commitments can be part of this normative standard. Similarly, when applied to collective agents, retrospective responsibility is also collective: the outcome can be explained in terms of the collective agent's motivations, intentions, goals and reasons.

Over the course of this thesis I have attempted to show that rationality requirements do not only govern intentional individual agency, but also social and collective agency. Normative requirements give rise to normative expectations which, in turn, are central to attributions of responsibility for past outcomes. I have insisted throughout the previous Chapters that this conception of normativity was not moral, but *rational*.

I hope to have shown that the framework presented in this dissertation is apt as a starting point for extending the investigation about the nature of practical commitments and the structure of social obligations. I will now sketch two possible lines of future research in order to bring this project to a close.

First, I have made an effort to keep moral normativity aside, for my interest was to focus on rationality. It would be fruitful to put them back together, namely, to include moral reasons and norms in the normative structure of agency. Social and collective action have an intrinsic moral dimension that I have analytically ignored in the conceptualization of rational requirements—in fact, one of the claims I have defended in this work is that this separation is something that can actually be done, at least conceptually. However, I am not unaware of the fact that, if the framework I have presented is to be empirically applied or tested, moral norms ought to be introduced. Not only it is rationally incorrect to break a promise: it is also immoral in many contexts. By the same token, offering an excuse entails a moral dimension that affects the explanatory role of the agent in the production of the outcome. Lastly, the evolutionarily shaped mechanisms that enable social commitments, explored in Chapter 3, may also include the evolution of morality as having a central role in the explanation of those mechanisms. Thus, moral considerations

ought to be reintroduced in the framework, specially through means of a naturalized characterization of moral reasons and norms, and their role as promoters of cooperation.

Second, I have focused here on *practical* commitments, i.e. commitments to act in order to achieve a goal. I have argued that these commitments are normative insofar there is a normative relation between reasons, judgements, intentions and actions—the so-called rationality requirements. I have intentionally let aside the epistemic dimension of both practical commitments and rationality requirements. On the one hand, a practical commitment is based on normative *reasons*, which are *believed* to be the case by the agent. Thus, there might—and probably must—be certain epistemic conditions that justify to believe in the correctness of a normative judgement given the reasons that are used as premises for reaching that conclusion. In fact, my account of practical reasoning, which states that the conclusion of a practical reasoning process is not an intention nor an action but a normative *belief*, relies on the normative requirements of *theoretical* rationality, which I presuppose, but do not develop. Further research ought to explain the relation between practical and epistemic commitments, on the one hand, and between the normative requirements of practical and theoretical rationality, on the other.

References

Álvarez, J. M.C. 2006. El análisis de las interacciones grupales: las aplicaciones SOCIOS. *Anuario de psicología* **37** (3).

Adams, F., and A.R. Mele. 1992. The Intention/Volition Debate. *Canadian Journal of Philosophy* **22** (3): 323–337.

Ainslie, G. 2001. *Breakdown of Will*. Cambridge University Press.

Alexander, Richard D. 1987. *The Biology of Moral Systems*. Transaction Publishers.

Alicke, M.D. 1992. Culpable Causation. *Journal of Personality and Social Psychology* **63** (3): 368.

Alvarez, M. 2009. Actions, thought-experiments and the 'Principle of alternate Possibilities'. *Australasian Journal of Philosophy* **87** (1): 61–81.

Alvarez, M. 2005. Agents, Actions and Reasons. *Philosophical Books* **46** (1): 45–58.

Alvarez, M. 2010a. *Kinds of Reasons: an Essay in the Philosophy of Action*. Oxford: Oxford University Press.

Alvarez, M. 2010b. Reasons for Action and Practical Reasoning. *Ratio* **23** (4): 355–373.

Anderson, J. 2008. Verantwortung. In *Handbuch der politischen Philosophie und Sozialphilosophie*, edited by S. Gosepath, W. Hinsch, and B. Rössler. Berlin: De Gruyter.

Andreou, C. 2006. Standards, Advice, and Practical Reason. *Journal of Moral Philosophy* **3** (1): 57–67.

Andreou, Chrisoula, and Mark D White. 2010. *The Thief of Time : Philosophical Essays on Procrastination*. New York: Oxford University Press.

Anscombe, G.E.M. 1957. *Intention*. Cambridge Mass.: Harvard University Press.

Attfield, Robin. 2009. Mediated Responsibilities, Global Warming, and the Scope of Ethics. *Journal of Social Philosophy* **40** (2): 225–236.

Audi, Robert. 1993. *Action, Intention, and Reason*. Ithaca, NY: Cornell University Press.

Audi, Robert. 1973. Intending. *The Journal of Philosophy* **70** (13): 387–403.

Audi, Robert. 2006. *Practical Reasoning and Ethical Decision*. London; New York: Routledge.

Audi, Robert. 2004. Reasons, Practical Reason, and Practical Reasoning. *Ratio* **17** (2): 119–149.

Audi, Robert. 2001. *The Architecture of Reason: the structure and substance of rationality*. Oxford University Press, USA.

Austin, J.L. 1956. A Plea for Excuses: The Presidential Address. In *Proceedings of the Aristotelian Society*, 57:1–30. Vol. 57.

Axelrod, R. 1980. Effective Choice in the Prisoner's Dilemma. *Journal of Conflict Resolution* **24** (1): 3.

Axelrod, R., and W. D. Hamilton. 1981. The Evolution of Cooperation. *Science* **211** (4489): 1390.

Back, I., and A. Flache. 2008. The Adaptive Rationality of Interpersonal Commitment. *Rationality and Society* **20** (1): 65.

Bacon, Francis. 1630. *The Elements of the Common Lawes of England*. 2003rd ed. The Lawbook Exchange, Ltd.

Balliet, D. 2010. Communication and Cooperation in Social Dilemmas: A Meta-analytic Review. *Journal of Conflict Resolution* **54** (1): 39.

Barbalet, J. 2009. A Characterization of Trust, and its Consequences. *Theory and Society* **38** (4): 367–382.

Baron, Marcia. 2006. Excuses, Excuses. *Criminal Law and Philosophy* **1** (1): 21–39.

Barros, D.B. 2011. Negative Causation in Causal and Mechanistic Explanation. *Synthese*: 1–21.

Battigalli, P., and M. Dufwenberg. 2007. Guilt in Games. *The American Economic Review* **97** (2): 170–176.

Baumeister, R.F., E. Bratslavsky, M. Muraven, and D.M. Tice. 1998. Ego Depletion: Is the Active Self a Limited Resource? *Journal of Personality and Social Psychology* **74** (5): 1252.

Baumeister, R.F., K.D. Vohs, and D.M. Tice. 2007. The Strength Model of Self-control. *Current directions in psychological science* **16** (6): 351–355.

Becker, G. S. 1974. A Theory of Social Interactions. *The Journal of Political Economy* **82** (6): 1063–1093.

Becker, H. S. 1960. Notes on the Concept of Commitment. *American Journal of Sociology* **66** (1): 32–40.

Beebee, H. 2004. Causing and Nothingness. In *Causation and counterfactuals*, edited by J. D Collins, E. J Hall, and L. A Paul, 291–308. The MIT Press.

van den Bergh, J. C.J.M, and J. M Gowdy. 2009. A group selection perspective on economic behavior, institutions and organizations. *Journal of Economic Behavior and Organization* **72** (1): 1–20.

Berman, M.N. 2003. Justification and Excuse, Law and Morality. *Duke Law Journal*: 1–77.

Bicchieri, C., and A. Chavez. 2010. Behaving as expected: Public information and fairness norms. *Journal of Behavioral Decision Making* **23** (2): 161–178.

Bicchieri, C., Julian Nida-Rümelin, and Wolfgang Spohn. 2000. Words and Deeds: A Focus Theory of Norms. In *Rationality, Rules and Structure*. Dordrecht: Kluwer Academic Publishers.

Bicchieri, C. 2006. *The Grammar of Society: The Emergence and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

Bicchieri, C., and E. Xiao. 2008. Do the Right Thing: But Only if Others Do so. *Journal of Behavioral Decision Making* **21**: 1–18.

Birnbacber, D. 2001. Philosophical Foundations of Responsibility. In *Responsibility: The many Faces of a Social Phenomenon*, edited by A.E. Auhagen and H.W. Bierhoff, 9. London: Routledge.

Björnsson, G., and Karl Persson. 2012. The Explanatory Component of Moral Responsibility. *Noûs* **46** (2): 326–354.

Boadway, R., Z. Song, and J. F Tremblay. 2007. Commitment and Matching Contributions to Public Goods. *Journal of Public Economics* **91** (9): 1664–1683.

Bok, H. 1998. *Freedom and Responsibility*. Princeton Univ Pr.

Boniolo, Giovanni, and Gabriele De Anna. 2006. The Four Faces of Omission -- Ontology, Terminology, Epistemology, and Ethics. *Philosophical Explorations* **9** (3): 277.

Bosman, R., and F. Van Winden. 2002. Emotional Hazard in a Power-to-take Experiment. *Economic Journal* **112** (476): 147–169.

Bouvier, A. 2007. Collective Belief, Acceptance, and Commitment in Science: The Copenhagen School Example. *Iyyun: The Jerusalem Philosophical Quarterly* **56**: 91–118.

Bouvier, A. 2004. Individual beliefs and collective beliefs in sciences and philosophy: The plural subject and the polyphonic subject accounts. *Philosophy of the social sciences* **34** (3): 382–407.

Boyd, Robert, Samuel Bowles, Peter J. Richerson, and H. Gintis. 2003. The Evolution of Altruistic Punishment. *Proceedings of the National Academy of Sciences of the United States of America* **100** (6): 3531–3535.

Braham, M., and M. van Hees. 2010. An Anatomy of Moral Responsibility. Manuscript. University of Gronigen.

Braham, M., and M. van Hees. 2009. Degrees of Causation. *Erkenntnis* **71** (3): 323–344.

Braham, M., and M. van Hees. 2011. Responsibility Voids. *The Philosophical Quarterly* **61** (242): 6–15.

Bramoullé, Y. 2007. Anti-coordination and Social Interactions. *Games and Economic Behavior* **58** (1): 30–49.

Brandom, R. 1998. Action, norms, and practical reasoning. *Noûs* **32** (S12): 127–139.

Brandom, R. 1983. Asserting. *Noûs* **17** (4): 637–650.

Brandom, R. 1994. *Making it Explicit*. Harvard University Press.

Bratman, M. 1987. *Intention, plans, and practical reason*. Cambridge Mass.: Harvard University Press.

Bratman, M. 2009a. Intention, Practical Rationality, and Self-Governance. *Ethics* **119** (3): 411–443.

Bratman, M. 2000. Reflection, Planning, and Temporally Extended Agency. *The Philosophical Review* **109** (1): 35–61.

Bratman, M. 1992. Shared cooperative activity. *The Philosophical Review*: 327–341.

Bratman, M. 1993. Shared intention. *Ethics* **104** (1): 97–113.

Bratman, M. 2004. Shared Valuing and Practical Reasoning. In *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, edited by R. Jay Wallace, 1–27.

Bratman, M. 1998. Toxin, temptation, and the stability of intention. *Rational Commitment and Social Justice. Essays for Gregory Kavka, Cambridge University Press, Cambridge*.

Bratman, M. 1984. Two faces of intention. *The Philosophical Review*: 375–405.

Bratman, M. 2001. Two Problems about Human Agency. *Proceedings of the Aristotelian Society* **101** (1): 309–326.

Bratman, M.E. 2009b. Shared Agency. In *Philosophy of the social sciences*, edited by C. Mantzavinos, 41–59. Cambridge: Cambridge University Press.

Brennan, G. 2007. The Grammar of Rationality. In *Rationality and Commitment*, edited by F. Peter and H. B. Schmid. Oxford: Oxford University Press.

Broome, J. 2001a. Are intentions reasons? And how should we cope with incommensurable values? In *Practical rationality and preference: Essays for David Gauthier*, edited by C.W. Morris and A. Ripstein, 98–120. Cambridge University Press, Cambridge.

Broome, J. 2001b. Normative Practical Reasoning: John Broome. In *Aristotelian Society Supplementary Volume*, 75:175–193. Vol. 75.

Broome, J. 1999. Normative requirements. *Ratio* **12** (4): 398–419.

Broome, J. 2002. Practical reasoning. In *Practical reasoning*, 85–111. Oxford: Oxford University Press.

Broome, J. 2010. Rationality. In *A Companion to the Philosophy of Action*, edited by O. Timothy and C. Sandis, 283–292. Wiley-Blackwell.

Broome, J. 2004. Reasons. *Reason and value: Themes from the moral philosophy of Joseph Raz*: 28–55.

Broome, J. 2007. Wide or Narrow Scope? *Mind* **116** (462): 359.

Brunero, J. 2010. The scope of rational requirements. *The Philosophical Quarterly* **60** (238): 28–49.

Bshary, R., and R. Bergmuller. 2008. Distinguishing four fundamental approaches to the evolution of helping. *Journal of evolutionary biology* **21** (2): 405–420.

Carson, T.L. 2006. The Definition of Lying. *Nous* **40** (2): 284–306.

Castelfranchi, C. 1995. Commitments: From Individual Intentions to Groups and Organizations. In *Proc. First Int. Conf. on Multi-Agent Systems (ICMAS-95)*, 41–48. San Francisco.

Castelfranchi, C., and R. Falcone. 2002. Social trust: A cognitive approach. *Trust and deception in virtual societies*: 55–90.

Castelfranchi, C., and R. Falcone. 1999. The dynamics of trust: From beliefs to action. In *Proceedings of the Second Workshop on Deception, Fraud and Trust in Agent Societies*, 41–54.

Castelfranchi, C., and M. Guerini. 2007. Is it a promise or a threat? *Pragmatics & Cognition* **15** (2): 277–311.

Castelfranchi, C., and F. Paglieri. 2007. The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions. *Synthese* **155** (2): 237–263.

Castelfranchi, C. 1999. Prescribed mental attitudes in goal-adoption and norm-adoption. *Artificial Intelligence and Law* **7** (1): 37–50.

Castelfranchi, C. 1998. Towards an Agent Ontology:Autonomy, Delegation, Adaptivity. *AI*IA Notizie*: pp.45–50.

Castelfranchi, C. 2008. Trust and reciprocity: misunderstandings. *International Review of Economics* **55** (1): 45–63.

Chant, S. R, and Z. Ernst. 2008. Epistemic conditions for collective action. *Mind* **117** (467): 549–573.

Charness, G., and M. Dufwenberg. 2010. Bare promises: An experiment. *Economics Letters* **107** (2): 281–283.

Charness, G., and M. Rabin. 2002. Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics* **117** (3): 817.

Clark, Herbert H. 1996. *Using language*. Cambridge University Press.

Clarke, Randolph. 2010. Intentional Omissions. *Noûs* **44** (1): 158–177.

Cohen, D., and J. Vandello. 2001. Honor and 'faking' honorability. *Evolution and the capacity for commitment*: 163–185.

Cohen, P. R, and H. J Levesque. 1990. Intention is choice with commitment. *Artificial intelligence* **42** (2-3): 213–261.

Copp, D. 2006. On the Agency of Certain Collective Entities: An Argument from 'Normative Autonomy'. *Midwest Studies In Philosophy* **30** (1): 194–221.

Copp, D. 2007. The collective moral autonomy thesis. *Journal of Social Philosophy* **38** (3): 369–388.

Cosmides, L., and J. Tooby. 2004. Evolutionary Psychology and the Emotions. *Handbook of emotions*: 91.

Croson, R., and J. Konow. 2009. Social preferences and moral biases. *Journal of Economic Behavior & Organization* **69** (3): 201–212.

Cullity, G. M. 2008. Decisions, Reasons and Rationality. *Ethics* **119** (1): 57–95.

Dancy, J. 2004a. *Ethics without principles*. Oxford University Press, USA.

Dancy, J. 2000. *Practical reality*. Oxford University Press, USA.

Dancy, J. 2004b. Two ways of explaining actions. *Royal Institute of Philosophy Supplement* **55** (-1): 25–42.

Daniel Cohen, and Toby Handfield. 2010. Rational Capacities, Resolve, and Weakness of Will. *Mind* **119** (476): 907–932.

Davidson, D. 1963. Actions, reasons, and causes. *The Journal of Philosophy* **60** (23): 685–700.

Davidson, D. 1967. Causal relations. *The Journal of Philosophy* **64** (21): 691–703.

Davidson, D. 1980. How Is Weakness of the Will Possible? In *Essays on actions and events*, 21–42. Oxford: Clarendon Press.

Davis, W.A. 1984. A causal theory of intending. *American Philosophical Quarterly* **21** (1): 43–54.

Debreu, G. 1959. *Theory of value: An axiomatic analysis of economic equilibrium*. Yale University Press.

Dodd, D. 2009. Weakness of Will as Intention-Violation. *European Journal of Philosophy* **17** (1): 45–59.

Dovidio, J. F, and L. A Penner. 2004. Helping and Altruism. In *Emotion and motivation*, edited by Marilynn B. Brewer and Miles Hewstone, 247. London: Blackwell Pub.

Dowe, P. 2001. A counterfactual theory of prevention and'Causation'by omission. *Australasian Journal of Philosophy* **79** (2): 216–226.

Dowe, P. 2009. Absences, possible causation, and the problem of non-locality. *The Monist* **92** (1): 23–40.

Dowe, P. 2004. Causes are physically connected to their effects: why preventers and omissions are not causes. In *Contemporary debates in philosophy of science*, edited by C.R. Hitchcock, 189–196. Malden MA: Blackwell.

Dowe, P. 2000. *Physical Causation*. Cambridge: Cambridge University Press.

Driver, J. 2008. Kinds of norms and legal causation: Reply to Knobe and Fraser and Deigh. *Moral psychology* **2**: 459–62.

Duff, A. 2009a. Legal and Moral Responsibility. *Philosophy Compass* **4** (6): 978–986.

Duff, R. A. 2009b. Strict Responsibility, Moral and Criminal. *The Journal of Value Inquiry* **43** (3): 295–313.

Duff, RA. 2007. Excuses, moral and legal: a comment on Marcia Baron's 'excuses, excuses'. *Criminal Law and Philosophy* **1** (1): 49–55.

Eccles, J.S., and A. Wigfield. 2002. Motivational beliefs, values, and goals. *Annual review of psychology* **53** (1): 109–132.

Ekstrom, L.W. 2005. Alienation, autonomy, and the self. *Midwest studies in philosophy* **29** (1): 45–67.

Ellingsen, T., M. Johannesson, S. Tjotta, and G. Torsvik. 2010. Testing guilt aversion. *Games and Economic Behavior* **68** (1): 95–107.

Elster, J. 2003. Don't Burn Your Bridge Before You Come to It: Some Ambiguities and Complexities of Precommitment. *Tex. L. Rev.* **81** (7): 1751.

Elster, J. 1979. *Ulysses and the Sirens. Studies in Rationality and Irrationality.*

Elster, J. 2000. *Ulysses unbound: Studies in rationality, precommitment, and constraints*. Cambridge Univ Pr.

Elster, J. 2006. Weakness of Will and Preference Reversals. In *Understanding choice, explaining behaviour : essays in honour of Ole-Jørgen Skog*, edited by Ole-Jørgen Skog and Jon Elster, 57–74. Oslo: Oslo Academic Press.

Enoch, D. 2006. Agency, Shmagency: Why Normativity Won't Come from What Is Constitutive of Action. *The Philosophical Review* **115** (2): 169.

Evans, J.S.B.T. 2008. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.* **59**: 255–278.

Everson, S. 2009. What are Reasons for Action? In *New essays on the explanation of action*, edited by C. Sandis. Palgrave Macmillan.

Fehr, E., U. Fischbacher, and S. Gächter. 2002. Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature* **13** (1): 1–25.

Fehr, E., and U. Fischbacher. 2004a. Social norms and human cooperation. *Trends in Cognitive Sciences* **8** (4): 185–190.

Fehr, E., and U. Fischbacher. 2003. The nature of human altruism. *Nature* **425** (6960): 785–791.

Fehr, E., and U. Fischbacher. 2004b. Third-party punishment and social norms. *Evolution and human behavior* **25** (2): 63–87.

Fehr, E., and S. Gächter. 2002. Altruistic punishment in humans. *Nature* **415** (6868): 137–140.

Fehr, E., and J. Henrich. 2003. Is strong reciprocity a maladaptation? On the evolutionary foundations of human altruism.

Fehr, E., and B. Rockenbach. 2004. Human altruism: economic, neural, and evolutionary perspectives. *Current Opinion in Neurobiology* **14** (6): 784–790.

Fehr, E., and K. M Schmidt. 1999. A theory of fairness, competition, and cooperation. *Quarterly journal of Economics* **114** (3): 817–868.

Ferrero, L. 2009. Constitutivism and the Inescapability of Agency. *Oxford Studies in Metaethics* **4**: 303–333.

Ferrero, L. 2006. Three ways of spilling ink tomorrow. In *Rationality in belief and action*, edited by E. Baccarini and s. Prijic-Samarzija, 95–127. Rijeka: Filozofski Facultet.

Finlay, S. 2006. The reasons that matter. *Australasian Journal of Philosophy* **84** (1): 1–20.

Fischer, J. M. 2006. *My way: Essays on moral responsibility*. Oxford University Press, USA.

Fischer, J. M, and M. Ravizza. 2000. *Responsibility and control: A theory of moral responsibility*. Cambridge Univ Pr.

Fischer, J. M. 1997. Responsibility, control, and omissions. *The Journal of Ethics* **1** (1): 45–64.

Fischer, J.M. 1998. Moral responsibility and the metaphysics of free will: Reply to Van Inwagen. *The Philosophical Quarterly* **48** (191): 215–220.

Fischer, J.M. 2010. Precis of My Way: Essays on Moral Responsibility. *Philosophy and Phenomenological Research* **80** (1): 229–241.

Fischer, J.M., and N.A. Tognazzini. 2009. The truth about tracing. *Nous* **43** (3): 531–556.

Fischer, John Martin. 2007. *Four views on free will*. John Wiley & Sons.

Van Fraassen, B.C. 1980. *The scientific image*. Oxford University Press, USA.

Frank, R. H. 2003. Commitment problems in the theory of rational choice. *Tex. L. Rev.* **81** (7): 1789.

Frank, R. H. 2001. Cooperation through emotional commitment. *Evolution and the capacity for commitment* **3**: 57–76.

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

Frankfurt, H.G. 1971. Freedom of the Will and the Concept of a Person. *The Journal of Philosophy* **68** (1): 5–20.

Frankfurt, H.G. 1988. *The importance of what we care about: philosophical essays*. Cambridge Univ Pr.

French, P. A. 1998. *Individual and collective responsibility*. Schenkman Books.

French, Peter A., Howard K. Wettstein, and John Martin Fischer. 2005. *Free will and moral responsibility*. Wiley-Blackwell.

Funkhouser, E. 2002. Three varieties of causal overdetermination. *Pacific philosophical quarterly* **83** (4): 335–351.

Güth, W., and H. Kliemt. 2004. The rationality of rational fools-The role of commitments, persons and agents in rational choice modeling. *Papers on Strategic Interaction*.

Gibbons, J. 2010. Things that make things reasonable. *Philosophy and Phenomenological Research* **81** (2): 335–361.

Gideon Yaffe. 2001. Recent Work on Addiction and Responsible Agency. *Philosophy & Public Affairs* **30** (2): 178–221.

Gigerenzer, G., and R. Selten. 2002. *Bounded Rationality: The Adaptive Toolbox*. MIT Press.

Gilbert, M. 2006. *A theory of political obligation: membership, commitment, and the bonds of society*. Oxford University Press, USA.

Gilbert, M. 1999. Obligation and Joint Commitment. *Utilitas* **11** (02): 143–163.

Gilbert, M. 1992. *On social facts*. Princeton Univ Pr.

Gilbert, M. 2009. Shared intention and personal intentions. *Philosophical studies* **144** (1): 167–187.

Gilbert, M. 2011. Three Dogmas about Promising. In *Promises and Agreements*, edited by H. Sheinman, 1:80–109. Vol. 1. Oxford: Oxford University Press.

Gilbert, M. 1990. Walking together: A paradigmatic social phenomenon. *Midwest Studies in Philosophy* **15** (1): 1–14.

Gilead, A. 1999. How is Akrasia possible after all? *Ratio* **12** (3): 257–270.

Gintis, H. 2000a. Group selection and human prosociality. *Journal of Consciousness Studies, 7* **1** (2): 215–219.

Gintis, H., E. A Smith, and S. Bowles. 2001. Costly signaling and cooperation. *Journal of theoretical biology* **213** (1): 103–119.

Gintis, H. 2003. Solving the puzzle of prosociality. *Rationality and Society* **15** (2): 155.

Gintis, H. 2000b. Strong reciprocity and human sociality. *Journal of Theoretical Biology* **206** (2): 169–179.

Gollwitzer, P.M. 1990. Action phases and mind-sets. In *Handbook of motivation and cognition: Foundations of social behavior*, edited by R.M. Sorrentino and E.T. Higgins, 2:53–92. Vol. 2. The Guilford Press.

Gollwitzer, P.M. 1993. Goal achievement: The role of intentions. *European review of social psychology* **4** (1): 141–185.

Gollwitzer, P.M. 1999. Implementation intentions. *American Psychologist* **54** (7): 493–503.

Gollwitzer, P.M., and P. Sheeran. 2006. Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in experimental social psychology* **38**: 69–119.

Gonzaga, G. C, D. Keltner, E. A Londahl, and M. D Smith. 2001. Love and the commitment problem in romantic relations and friendship. *Journal of Personality and Social Psychology* **81** (2): 247–262.

Gosling, J. C.B. 1990. *Weakness of the Will*. Routledge.

Graham, K. 2002. *Practical reasoning in a social world: how we act together*. Cambridge: Cambridge University Press.

Greene, J., and J. Cohen. 2004. For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society B: Biological Sciences* **359** (1451): 1775.

Guerini, M., and C. Castelfranchi. 2007. Promises and threats in persuasion. *Pragmatics and Cognition* **15** (2): 277–311.

Gur, N. 2011. Are Legal Rules Content-Independent Reasons? *Problema* **5**: 175–210.

Habib, Allen. 2008. Promises. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Winter 2008. Available from <http://plato.stanford.edu/archives/win2008/entries/promises/>. . Accessed2 April 2012.

Hall, N. 2004. Two concepts of causation. In *Causation and counterfactuals*, edited by J. D Collins and N. Hall, 225–276.

Hamilton, W. D. 1964. The genetical evolution of social behavior, parts 1 and 2. *Journal of Theoretical Biology* **7**: 1–52.

Hammerstein, Peter. 2003. *Genetic and cultural evolution of cooperation*. MIT Press.

Hardin, R. 2003. Gaming trust. In *Trust and reciprocity: Interdisciplinary lessons from experimental research*, edited by E. Ostrom and J. Walker, 80–101. New York: Russell Sage Foundation Publications.

Harman, E. 2009. 'I'll Be Glad I Did It' Reasoning and the Significance of Future Desires. *Philosophical Perspectives* **23** (1): 177–199.

Hart, H. L.A. 1982. *Essays on Bentham: studies in jurisprudence and political theory*. Oxford: Oxford University Press.

Hart, H.L.A., and John Gardner. 2008. *Punishment and Responsibility*. Oxford University Press.

Hart, H.L.A., and T. Honoré. 1985. *Causation in the Law*. Clarendon Press Oxford.

Hausman, D. M. 2005. Sympathy, Commitment, And Preference. *Economics and Philosophy* **21** (01): 33–50.

Heatherton, Todd F. 2011. Neuroscience of Self and Self-Regulation. *Annual review of psychology* **62**: 363–390.

Heatherton, Todd F., and K.D. Vohs. 1998. Why is it so difficult to inhibit behavior? *Psychological Inquiry* **9** (3): 212–216.

Held, V. 1970. Can a random collection of individuals be morally responsible? *The Journal of Philosophy* **67** (14): 471–481.

Henden, E. 2008. What is self-control? *Philosophical Psychology* **21** (1): 69–90.

Henrich, J., and F. J Gil-White. 2001. The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior* **22** (3): 165–196.

Hinchman, E.S. 2009. Receptivity and the Will. *Noûs* **43** (3): 395–427.

Hindriks, F. 2009. Corporate Responsibility and Judgment Aggregation. *Economics and Philosophy* **25** (02): 161–177.

Hirshleifer, J. 2001. On the emotions as guarantors of threats and promises. In *The dark side of the force: economic foundations of conflict theory*, 198. Cambridge University Press.

Hitchcock, C.R. 1996. The role of contrast in causal and explanatory claims. *Synthese* **107** (3): 395–419.

Hoekstra, Rinke, and Joost Breuker. 2007. Commonsense Causal Explanation in a Legal Domain. *Artificial Intelligence and Law* **15** (3): 281–299.

Hoffman, E., K. McCabe, K. Shachat, and V. Smith. 1994. Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior* **7** (3): 346–380.

Holton, R. 2003. How is Strength of Will Possible? In *Weakness of will and practical irrationality*, edited by Sarah Stroud and C. Tappolet, 39–67. Oxford, New York: Oxford University Press, USA.

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

Holton, R. 2004. Rational resolve. *The Philosophical Review* **113** (4): 507–535.

Holton, R., and S. Shute. 2007. Self-control in the modern provocation defence. *Oxford Journal of Legal Studies* **27** (1): 49–73.

Holton, R. 2006. The act of choice. *Philosophers' Imprint* **6** (3): 1–15.

Holton, R. 2009. *Willing, wanting, waiting*. Oxford University Press, USA.

Hume, D. 2007. *A treatise of human nature [1739]*. Vol. 1. Oxford University Press, USA.

Hussain, N. 2007. The Requirements of Rationality.

Van Inwagen, P. 1978. Ability and responsibility. *Philosophical Review* **87** (2): 201–204.

Irons, W. 2001. Religion as a hard-to-fake sign of commitment. *Evolution and the capacity for commitment*: 292–309.

Isaacs, Tracy Lynn. 2011. *Moral Responsibility in Collective Contexts*. Oxford University Press.

Job, V., C.S. Dweck, and G.M. Walton. 2010. Ego Depletion—Is It All in Your Head? *Psychological science* **21** (11): 1686.

Kahneman, D. 2003. A Perspective on Judgment and Choice: Mapping Bounded Rationality. *American Psychologist* **58** (9): 24.

Kalis, A., A. Mojzisch, T. S Schweizer, and S. Kaiser. 2008. Weakness of will, akrasia, and the neuropsychiatry of decision making: An interdisciplinary perspective. *Cognitive, Affective, & Behavioral Neuroscience* **8** (4): 402–417.

Kane, R. 2011. *The Oxford handbook of free will*. Oxford Univ Pr.

Kaufman, A.S. 1966. Practical decision. *Mind* **75** (297): 25–44.

Kavka, G. S. 1983. The toxin puzzle. *Analysis* **43** (1): 33.

Kearns, S., and D. Star. 2009. Reasons as Evidence. *Oxford Studies in Metaethics: Volume Four*: 215.

Keating, G.C. 1996. Reasonableness and rationality in negligence theory. *Stanford Law Review* **48**: 311.

Keltner, D., J. Haidt, and M. N Shiota. 2006. Social Functionalism and the Evolution of Emotions. In *Evolution and Social Psychology*, edited by Mark Schaller, Jeffry A. Simpson, and Kenrick, Douglas T., 115. Frontiers of Social Psychology. Psychology Press.

Kerr, N. L, J. Garst, D. A Lewandowski, and S. E Harris. 1997. That still, small voice: Commitment to cooperate as an internalized versus a social norm. *Personality and social psychology Bulletin* **23** (12): 1300.

Kerr, N. L, and C. M Kaufman-Gilliland. 1994. Communication, commitment, and cooperation in social dilemma. *Journal of Personality and Social Psychology* **66** (3): 513–529.

Klass, G. 2009. A Conditional Intent to Perform. *Legal Theory* **15** (02): 107–147.

Klein, G.A. 1999. *Sources of power: How people make decisions*. The MIT Press.

Knobe, J. 2006. Folk psychology, folk morality. Princeton University.

Knobe, J., and B. Fraser. 2008. Causal judgment and moral judgment: Two experiments. *Moral psychology* **2**: 441–8.

Kolodny, N. 2005. Why be rational? *Mind* **114** (455): 509–563.

Kolodny, Niko, and R. Jay Wallace. 2003. Promises and Practices Revisited. *Philosophy and Public Affairs* **31** (2): 119–154.

Korsgaard, Christine Marion, and Onora O'Neill. 1996. *The sources of normativity*. Cambridge University Press.

Krebs, J. R, and R. Dawkins. 1984. Animal signals: mind-reading and manipulation. *Behavioural Ecology: an evolutionary approach* **2**: 380–402.

Kurzban, R., K. McCabe, V. L Smith, and B. J Wilson. 2001. Incremental commitment and reciprocity in a real-time public goods game. *Personality and Social Psychology Bulletin* **27** (12): 1662.

Kutz, C. 2000. *Complicity: Ethics and law for a collective age*. Cambridge Univ Pr.

Langton, Rae. 2004. Intention as Faith. In *Agency and action*, edited by John Hyman, 243–258. New York: Cambridge University Press.

Larvor, B. 2010. Frankfurt counter-example defused. *Analysis* **70** (3): 506.

Lemaire, Stéphane. 2012. A Gate-Based Account of Intentions. *Dialectica*. Available from <http://onlinelibrary.wiley.com/doi/10.1111/j.1746-8361.2011.01287.x/abstract>. . Accessed7 February 2012.

Lenk, H., and M. Maring. 1991. A Pie-Model of Moral Responsibility. In *Advances in Scientific Philosophy*, edited by G. Schurz and G.J.W. Dorn, 483–494. Amsterdam - Atlanta: Rodopi.

Levy, N. 2011. Resisting Weakness of the Will. *Philosophy and Phenomenological Research* **82** (1): 134–155.

Lewis, D. 1973. Causation. *The Journal of Philosophy* **70** (17): 556–567.

List, C., and P. Pettit. 2002. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy* **18** (1): 89–110.

List, C., and P. Pettit. 2004. Aggregating Sets of Judgments: Two Impossibility Results Compared 1. *Synthese* **140** (1): 207–235.

List, C., and P. Pettit. 2006. Group agency and supervenience. *The Southern journal of philosophy* **44** (S1): 85–105.

List, C., and P. Pettit. 2011. *Group Agency: The Possibility, Design and Status of Corporate Agents*. Oxford University Press.

List, C., and C. Puppe. 2009. Judgment aggregation: A survey. In *Handbook of Rational and Social Choice*, edited by P. Anand and P. K. Pattanaik, 457–483. Oxford; New York: Oxford University Press.

Longworth, F. 2006. Causation, Pluralism and Responsibility. *Philosophica* **77**: 45–68.

Lord, E. 2011. Violating Requirements, Exiting from Requirements, and the Scope of Rationality. *The Philosophical Quarterly* **61** (243): 392–399.

Lorini, E., and C. Castelfranchi. 2004. *To Attempt and To Try: for a cognitive theory of Action*. MindRACES: from reactive to anticipatory cognitive embodied systems.

Mason, E. 2005. We make no promises. *Philosophical studies* **123** (1): 33–46.

Mathiesen, K. 2006. We're All in This Together: Responsibility of Collective Agents and Their Members. *Midwest Studies In Philosophy* **30** (1): 240–255.

May, J., and R. Holton. 2011. What in the world is weakness of will? *Philosophical Studies*: 1–20.

May, L. 1990. Collective Inaction and Shared Responsibility. *Nous* **24** (2): 269–278.

May, Larry. 1992. *Sharing Responsibility*. University of Chicago Press.

McAdams, R.H., and E. Rasmusen. 2006. Norms in law and economics. In *Handbook of Law and Economics*, edited by A.M. Polinsky and S. Shavell, 2:1575–1612. Vol. 2. Amsterdam: North-Holland.

McGrath, S. 2005. Causation by omission: A dilemma. *Philosophical Studies* **123** (1): 125–148.

McIntyre, A. 2006. What is wrong with weakness of will? *Journal of philosophy* **103** (6): 284–311.

Mele, A. R. 2009. *Effective intentions: The power of conscious will*. Oxford University Press, USA.

Mele, A. R. 1987. *Irrationality: an essay on akrasia, self-deception, and self-control*. Oxford University Press, USA.

Mele, A. R. 2003a. *Motivation and agency*. Oxford University Press, USA.

Mele, A. R. 1992. *Springs of action*. Oxford University Press New York.

Mele, A. R. 2010. Weakness of will and akrasia. *Philosophical studies*: 1–14.

Mele, A.R. 1995. *Autonomous agents: From self-control to autonomy*. Oxford University Press, USA.

Mele, A.R. 2003b. Intending and Trying: Tuomela vs. Bratman at the video arcade. In *Realism in action*, edited by Matti Sintonen, Petri Ylikoski, and Kaarlo Miller, 129–136. Springer.

Mellema, G. 1985. Groups, Responsibility, and the Failure to Act. *International Journal of Applied Philosophy* **1985**: 57–66.

Mellema, G. 2003. Responsibility, taint, and ethical distance in business ethics. *Journal of Business Ethics* **47** (2): 125–132.

Menzies, P. 2004. Difference-making in context. edited by Collins.

Miller, K. 2003. Commitments. In *Realism in Action*, edited by Matti Sintonen, Petri Ylikoski, and Kaarlo Miller, 169–178. Springer.

Miller, K. 2006. Social obligation as reason for action. *Cognitive Systems Research* **7** (2-3): 273–285.

Moore, Michael S. 2009. *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*. Oxford University Press.

Morin, O. 2009. Y at-il des règles constitutives? *Tracés* (2): 109–125.

Mosquera, P. M.R, A. S.R Manstead, and A. H Fischer. 2002. The role of honour concerns in emotional reactions to offences. *Cognition & Emotion* **16** (1): 143–163.

Nagel, T. 1970. *The possibility of altruism*. Clarendon P.

Nelkin, Dana K. 2008. Responsibility and Pational Abilities: Defending an Asymmetrical View. *Pacific Philosophical Quarterly* **89** (4): 497–515.

Nesse, R. M. 2001. Natural selection and the capacity for subjective commitment. *Evolution and the Capacity for Commitment*: 1–44.

Nesse, R. M. 1994. Why is group selection such a problem? *Behavioral and Brain Sciences* **17** (04): 633–634.

Nichols, S., and J. Knobe. 2007. Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous* **41** (4): 663–685.

Nowak, M. A. 2006. Five rules for the evolution of cooperation. *Science* **314** (5805): 1560.

Nowak, M. A, and K. Sigmund. 1993. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner$\backslash$'s Dilemma game. *Nature* **364** (6432): 56–58.

Nowak, M. A, and K. Sigmund. 2005. Evolution of indirect reciprocity. *Nature* **437** (7063): 1291–1298.

O'Connor, T. 1996. Why agent causation? *Philosophical Topics* **24** (2): 143–58.

Origgi, G. 2008. *Qu'est-ce que la confiance?* J. Vrin.

Oshana, M.A.L. 1997. Ascriptions of responsibility. *American Philosophical Quarterly* **34** (1): 71–83.

Ostrom, E., J. Walker, and R. Gardner. 1992. Covenants with and without a sword: Self-governance is possible. *The American Political Science Review* **86** (2): 404–417.

Owens, D. 2006. A simple theory of promising. *The Philosophical Review* **115** (1): 51–77.

Pacherie, E. 2011. Self-Agency. In *The Oxford handbook of the self*, edited by Shaun Gallagher, 442–464. Oxford Handbooks Online.

Pacherie, E. 2008. The phenomenology of action: A conceptual framework. *Cognition* **107** (1): 179–217.

Parfit, D., and J. Broome. 1997. Reasons and motivation. *Proceedings of the Aristotelian society, supplementary volumes* **71**: 99–146.

Parfit, D. 2011. *On what matters*. Vol. 1. Oxford: Oxford University Press.

Parfit, D. 2001. Rationality and reasons. *Exploring practical philosophy: From action to values*: 17–39.

Parfit, D. 1984. *Reasons and persons*. Oxford. Oxford University Press.

Parisi, F. 2000. The cost of the game: a taxonomy of social interactions. *European Journal of Law and Economics* **9** (2): 99–114.

Paul, Sarah K. 2009. Intention, Belief, and Wishful Thinking: Setiya on 'Practical Knowledge'. *Ethics* **119** (3): 546–557.

Penner, L. A, J. F Dovidio, J. A Piliavin, and D. A Schroeder. 2005. Prosocial Behavior: Multilevel Perspectives. *Annual Review of Psychology*.

Petersson, Björn. 2008. Collective Omissions and Responsibility. *Philosophical Papers* **37** (2): 243.

Pettit, P. 2003a. Akrasia, collective and individual. In *Weakness of Will and Practical Irrationality*, edited by S. Stroud and C. Tappolet, 68–97. Oxford, New York: Oxford University Press.

Pettit, P. 2005. Construing Sen on commitment. *Economics and Philosophy* **21** (01): 15–32.

Pettit, P. 2001. Deliberative democracy and the discursive dilemma. *Philosophical Issues* **11** (1): 268–299.

Pettit, P. 2003b. Groups with Minds of their Own. In *Socializing Metaphysics*, edited by F. Schmitt, 167–93. Rowman and Littlefield.

Pettit, P. 2007a. Rationality, reasoning and group agency. *dialectica* **61** (4): 495–519.

Pettit, P. 2007b. Responsibility incorporated. *Ethics* **117**: 171–201.

Piller, C. 2005. Kinds of Practical Reasons: Attitude-Related Reasons and Exclusionary Reasons. *Coordenação*: 98.

van de Poel, I. 2011. The relation between forward-looking and backward-looking responsibility. *Moral Responsibility: Beyond Free Will and Determinism* **27**: 37.

Posner, R. A, and E. B Rasmusen. 1999. Creating and Enforcing Norms, with Special Reference to Sanctions. *International Review of Law and Economics* **19**: 369–382.

Pratt, M. G. 2007. Promises, Contracts and Voluntary Obligations. *Law and Philosophy* **26** (6): 531–574.

Pratt, M. 2002. Promises and perlocutions. *Critical Review of International Social and Political Philosophy* **5** (2): 93–119.

Raffoul, F. 2010. *The origins of responsibility*. Indiana Univ Pr.

Rawls, J. 1999. *A Theory of Justice*. Cambridge: Harvard University Press.

Rawls, J. 1955. Two concepts of rules. *The Philosophical Review* **64** (1): 3–32.

Raz, J. 1975. *Practical reason and norms*. Oxford University Press, USA.

Raz, J. 2001. Reasoning with Rules. *Current Legal Problems* **54** (1): 1–18.
Raz, J. 2009. Reasons: Explanatory and Normative. In *New Essays on the Explanation of Action*, edited by C. Sandis, 184–202. Palgrave Macmillan.
Raz, J. 1986. *The morality of freedom*. Oxford University Press, USA.
Raz, J. 2005. The myth of instrumental rationality. *Journal of Ethics and Social Philosophy* **1** (1): 2–28.
Rescorla, M. 2009. Assertion and its constitutive norms. *Philosophy and Phenomenological Research* **79** (1): 98–130.
Richardson, H. S. 1999. Institutionally divided moral responsibility. *Social Philosophy and Policy* **16** (02): 218–249.
Ridge, M. 1998. Humean intentions. *American Philosophical Quarterly* **35** (2): 157–178.
Rivera-Lopez, E. 2006. Promises, Expectations, and Rights. *Chi.-Kent L. Rev.* **81**: 21.
Robinson, Paul H. 1982. Criminal Law Defenses: A Systematic Analysis. *Columbia Law Review* **82** (2): 199–291.
Roskies, A.L. 2010. How does neuroscience affect our conception of volition? *Annual review of neuroscience* **33**: 109–130.
Ross, A. 1975. *On guilt, responsibility, and punishment*. Univ of California Pr.
Roxborough, C., and J. Cumby. 2009. Folk psychological concepts: Causation. *Philosophical Psychology* **22** (2): 205–213.
Royakkers, L., and F. P. M. Dignum. 1998. Collective obligation and commitment. In *Proceedings of 5th Int. conference on Law in the Information Society*.
Sánchez-Cuenca, I. 1998. Institutional commitments and democracy. *European Journal of Sociology* **39** (01): 78–109.
Salmon, W.C. 1998. *Causality and Explanation*. Oxford: Oxford University Press.
Sartorio, C. 2007. Causation and Responsibility. *Philosophy Compass* **2** (5): 749–765.
Sartorio, C. 2006. Disjunctive causes. *The Journal of philosophy* **103** (10): 521–538.
Sartorio, C. 2004. How To Be Responsible For Something Without Causing It. *Philosophical Perspectives* **18** (1): 315–336.
Sartorio, C. 2009. Omissions and Causalism. *Noûs* **43** (3): 513–530.
Scanlon, T. 1990. Promises and practices. *Philosophy & Public Affairs* **19** (3): 199–226.
Scanlon, T. 2003. *The difficulty of tolerance: essays in political philosophy*. Cambridge Univ Pr.
Scanlon, T. 1998. *What we owe to each other*. Belknap Press.
Schaffer, J. forthcoming. Causal Contextualisms. In *Contrastivism in Philosophy*, edited by M. Blaauw. Routledge.
Schaffer, J. 2005. Contrastive causation. *The Philosophical Review* **114** (3): 327–358.
Schaffer, J. 2010. Contrastive Causation in the Law. *Legal Theory* **16** (04): 259–297.
Schechtman, M. 2004. Self-expression and self-control. *Ratio* **17** (4): 409–427.
Schelling, Thomas C. 1956. An Essay on Bargaining. *The American Economic Review* **46** (3): 281–306.
Schelling, Thomas C. 2007. *Strategies of Commitment and Other Essays*. Harvard University Press.
Schelling, Thomas C. 1960. *The strategy of conflict*. Harvard University Press.
Schlesinger, H. J. 2008. *Promises, oaths, and vows: on the psychology of promising*. The Analytic Press.
Schmid, H. B. 2009. *Plural Action: Essays in Philosophy and Social Science*. Springer Verlag.
Schroeder, M. 2008. Having reasons. *Philosophical Studies* **139** (1): 57–71.

Schroeder, M. 2010. Knowledge is belief for sufficient (objective and subjective) reason. Manuscript. Available from <http://www-rcf.usc.edu/~maschroe/research/Schroeder_Knowledge_Is.pdf>.

Schroeder, M. 2009. Means-end coherence, stringency, and subjective reasons. *Philosophical studies* **143** (2): 223–248.

Schroeder, M. 2007. Reasons and agent-neutrality. *Philosophical Studies* **135** (2): 279–306.

Schroeder, T. 2004. *Three faces of desire*. Oxford University Press, USA.

Schueler, G. F. 1995. *Desire: its role in practical reason and the explanation of action*. MIT Press.

Schueler, G. F. 2003. *Reasons and purposes*. Clarendon Press.

Schweder, R. 1999. Causal explanation and explanatory selection. *Synthese* **120** (1): 115–124.

Searle, J. R. 1990. Collective intentions and actions. *Intentions in communication*: 401–415.

Searle, J.R. 1983. *Intentionality, an essay in the philosophy of mind*. Cambridge Univ Pr.

Searle, J.R. 2001. *Rationality in action*. MIT Press.

Searle, J.R. 2007. Social Ontology: The Problem and Steps toward a Solution. In *Intentional acts and institutional facts essays on John Searle's social ontology*, edited by Savas L Tsohatzidis, 11–30. Dordrecht, the Netherlands: Springer. Available from <http://site.ebrary.com/id/10187007>. . Accessed27 April 2012.

Sekhar Sripada, C. 2010. Philosophical Questions about the Nature of Willpower. *Philosophy Compass* **5** (9): 793–805.

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

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

Sen, A. 2005. Why Exactly Is Commitment Important for Rationality? *Economics and Philosophy* **21** (01): 5–14.

Sethi, R., and E. Somanathan. 2005. Norm Compliance and Strong Reciprocity. In *Moral sentiments and material interests: the foundations of cooperation in economic life*, edited by H. Gintis, S. Bowles, R. T Boyd, and E. Fehr, 229–250. The MIT Press.

Setiya, K. 2004. Against internalism. *Noûs* **38** (2): 266–298.

Setiya, K. 2007a. Cognitivism about Instrumental Reason. *Ethics* **117** (4): 649–673.

Setiya, K. 2007b. *Reasons without rationalism*. Princeton Univ Pr.

Sheinman, H. 2011. Introduction: Promises and Agreements. In *Promises and Agreements*, 3–58. Oxford: Oxford University Press.

Shpall, S. forthcoming. Wide and narrow scope. *Philosophical Studies*: 1–20.

Sidgwick, Henry. 1907. *The methods of ethics*. Hackett Publishing.

Silver, D. 2002. Collective Responsibility and the Ownership of Actions. *Public Affairs Quarterly*: 287–304.

Simon, H. A. 1990. A mechanism for social selection and successful altruism. *Science* **250** (4988): 1665.

Simon, H. A. 1993. Altruism and economics. *The American Economic Review* **83** (2): 156–161.

Simpson, J. A. 2007. Psychological foundations of trust. *Current directions in psychological science* **16** (5): 264.

Singh, M. P. 1999. An ontology for commitments in multiagent systems. *Artificial Intelligence and Law* **7** (1): 97–113.

Skorupski, J. 2002. The ontology of reasons. *Topoi* **21** (1): 113–124.

Smith, A. M. 2007. On being responsible and holding responsible. *The Journal of Ethics* **11** (4): 465–484.

Smith, John Maynard, and David Harper. 2003. *Animal signals*. Oxford University Press.

Smith, M. J., and D. G. C. Harper. 1995. Animal Signals: Models and Terminology. *Journal of Theoretical Biology* **177** (3): 305–311.

Smith, M. 1987. The Humean theory of motivation. *Mind* **96** (381): 36.

Smith, M. 1995. *The moral problem*. Blackwell.

Sneddon, A. 2006. *Action and responsibility*. Kluwer Academic Pub.

Sobel, Jordan Howard. 1994. *Taking chances: essays on rational choice*. Cambridge University Press.

Stanovich, K.E., and R.F. West. 2000. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and brain sciences* **23** (5): 645–665.

Sterelny, K., and B. Jeffares. 2010. Rational Agency in Evolutionary Perspective. In *A Companion to the Philosophy of Action*, edited by O. Timothy and C. Sandis, 374–383. Wiley-Blackwell.

Stilz, Anna. 2011. Collective Responsibility and the State. *Journal of Political Philosophy* **19** (2): 190–208.

Stout, R. 2004. XI—Internalising Practical Reasons. In *Proceedings of the Aristotelian Society (Hardback)*, 104:231–245. Vol. 104.

Strawson, F. 1962. Freedom and Resentment. 187: Vol. 187.

Streumer, B. 2010. Practical Reasoning. In *A Companion to the Philosophy of Action*, edited by O. Timothy and C. Sandis, 244–251. Wiley-Blackwell.

Strevens, M. 2004. The causal and unification approaches to explanation unified—causally. *Noûs* **38** (1): 154–176.

Stroud, Sarah. 2003. Weakness of Will and Practical Judgement. In *Weakness of Will and Practical Irrationality*, edited by Sarah Stroud and C. Tappolet, 121–146. Oxford, New York: Oxford University Press.

Sugden, R. 1998. Normative expectations: the simultaneous evolution of institutions and norms. In *Economics, values, and organization*, edited by Avner Ben-Ner and Louis G. Putterman, 73–100. Cambridge University Press.

Sugden, R. 2000a. Team preferences. *Economics and Philosophy* **16** (02): 175–204.

Sugden, R. 2000b. The motivating power of expectations. *Rationality, Rules and Structure*: 103–129.

Sugden, R. 1993. Thinking as a team: Towards an explanation of nonselfish behavior. *Social Philosophy and Policy* **10** (01): 69–89.

Tännsjö, T. 2007. The Myth of Innocence: On Collective Responsibility and Collective Punishment. *Philosophical Papers* **36** (2): 295–314.

Tenenbaum, S. 2010. Akrasia and Irrationality. In *A Companion to the Philosophy of Action*, edited by O. Timothy and C. Sandis, 274–281. Wiley-Blackwell.

Tenenbaum, S. 2007. The Conclusion of Practical Reason. *Moral psychology*: 323.

Thero, D. P. 2006. *Understanding moral weakness*. Vol. 183. Rodopi Bv Editions.

Thomson, J. J. 1990. *The realm of rights*. Harvard Univ Pr.

Thomson, J.J. 2007. Normativity. In *Oxford studies in metaethics*, edited by R. Shafer-Landau, 2: Vol. 2. Oxford University Press, USA.

Tognazzini, N. A. 2007. The Hybrid Nature of Promissory Obligation. *Philosophy and Public Affairs* **35** (3): 203.

Tollefsen, D. 2006. The Rationality of Collective Guilt. *Midwest Studies In Philosophy* **30** (1): 222–239.

van der Torre, L., and Y. Tan. 1999. Rights, Duties and Commitments between Agents. In *In Proceedings of the IJCAI'99*.

Trivers, R. L. 1971. The evolution of reciprocal altruism. *The Quarterly Review of Biology* **46** (1).

Tummolini, L., G. Andrighetto, C. Castelfranchi, and R. Conte. forthcoming. A convention or (tacit) agreements betwixt us: on reliance and its normative consequences. *Synthese*.

Tuomela, R., and W. Balzer. 1998. Collective acceptance and collective social notions. *Synthese* **117** (2): 175–205.

Tuomela, R. 2000. Belief versus acceptance. *Philosophical Explorations* **3** (2): 122–137.

Tuomela, R. 2002. *The philosophy of social practices: A collective acceptance view*. Cambridge Univ Pr.

Tuomela, R. 2007. *The philosophy of sociality: the shared point of view*. Oxford Univ Pr.

Vallentyne, P. 2006. Natural Rights and Two Conceptions of Promising. *The Chicago-Kent Law Review* **81** (1): 9–19.

Vallentyne, P. 2009. Responsibility and Compensation Rights. In *Hillel Steiner and the anatomy of justice: themes and challenges*, edited by Stephen De Wijze, Matthew H. Kramer, and Ian Carter, 16:85–98. Vol. 16. Routledge.

Vanberg, C. 2008. Why Do People Keep Their Promises? An Experimental Test of Two Explanations1. *Econometrica* **76** (6): 1467–1480.

Vargas, M. 2005. The trouble with tracing. *Midwest Studies in Philosophy* **29** (1): 269–291.

Velleman, J.D. 1997. Deciding How to Decide. In *Ethics and practical reason*, edited by G. Cullity and B.N. Gaut, 29–52. Oxford, New York: Oxford University Press, USA.

Velleman, J.D. 1989. *Practical reflection*. Princeton N.J.: Princeton University Press.

Velleman, J.D. 2000. *The possibility of practical reason*. Oxford University Press, USA.

Velleman, J.D. 1992. What happens when someone acts? *Mind* **101** (403): 461–481.

Vihvelin, K. 2008. Foreknowledge, Frankfurt, and Ability to Do Otherwise: A Reply to Fischer. *Canadian Journal of Philosophy* **38** (3): 343–372.

Vincent, N.A. 2011. A structured taxonomy of responsibility concepts. In *Moral responsibility: Beyond free will and determinism*, edited by N.A. Vincent, I. van de Poel, and J. van den Hoven. New York: Springer.

Vitek, William. 1993. *Promising*. Temple University Press.

Wallace, R. J. 2001. Normativity, commitment, and instrumental reason. *Philosophers' Imprint* **1** (3): 1–26.

Wallace, R. J. 1994. *Responsibility and the moral sentiments*. Harvard Univ Pr.

Wallace, R.J. 1999a. Addiction as defect of the will: Some philosophical reflections. *Law and Philosophy* **18** (6): 621–654.

Wallace, R.J. 1999b. Three conceptions of rational agency. *Ethical Theory and Moral Practice* **2** (3): 217–242.

Walton, D.N., and E.C.W. Krabbe. 1995. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. State Univ of New York Pr.

Watson, G. 2004a. *Agency and answerability*. Oxford, New York: Oxford University Press.

Watson, G. 2004b. Asserting and promising. *Philosophical Studies* **117** (1): 57–77.

Watson, G. 2004c. Disordered Appetites: Addiction, Compulsion, and Dependence. In *Agency and Answerability*, 59–88. Oxford, New York: Oxford University Press.

Watson, G. 1975. Free agency. *The Journal of Philosophy* **72** (8): 205–220.
Watson, G. 2009. Promises, Reasons and Normative Powers. In *Reasons for Action*, edited by D. Sobel and S. Wall, 155–178. Cambridge, Mass: Cambridge University Press.
Watson, G. 2005. Promising, Assurance and Expectation. Manuscript.
Watson, G. 1977. Skepticism about weakness of will. *The Philosophical Review* **86** (3): 316–339.
Watson, G. 2003. The Work of the Will. In *Weakness of will and practical irrationality*, edited by S. Stroud and C. Tappolet, 172–200. Oxford, New York: Oxford University Press, USA.
Watson, G. 1996. Two Faces of Responsibility. *Philosophical Topics* **24**: 227–248.
Way, J. 2010. Defending the wide-scope approach to instrumental reason. *Philosophical studies* **147** (2): 213–233.
Way, J. 2011. The symmetry of rational requirements. *Philosophical studies*: 1–13.
Way, J. 2009. Two Accounts of the Normativity of Rationality. *Journal of Ethics and Social Philosophy*.
Wedgwood, R. 2007a. Normativism defended. In *Contemporary Debates in the Philosophy of Mind*, edited by M. McLaughlin and J. Cohen, 85–102. Blackwell Pub.
Wedgwood, R. 2007b. *The nature of normativity*. Oxford University Press, USA.
Weiner, B. 1995. *Judgments of responsibility: A foundation for a theory of social conduct*. The Guilford Press.
West, S. A, A. S Griffin, and A. Gardner. 2008. Social semantics: how useful has group selection been? *Journal of Evolutionary Biology* **21** (1): 374–385.
West, S. A., A. S. Griffin, and A. Gardner. 2007. Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology* **20** (2): 415–432.
Westen, Peter. 2006. An Attitudinal Theory of Excuse. *Law and Philosophy* **25** (3): 289–375.
Westlund, A. C. 2009. Deciding Together. *Philosophical Imprint* **9**: 1–17.
Widerker, David, and Michael McKenna. 2006. *Moral responsibility and alternative possibilities: essays on the importance of alternative possibilities*. Ashgate Publishing, Ltd.
Wiley, R. H. 1983. The evolution of communication: information and manipulation. *Animal behaviour* **2**: 156–189.
Williams, B. 1982. *Moral luck: philosophical papers, 1973-1980*. Cambridge Univ Pr.
van Willigenburg, T. 2003. 7. Sources of normativity: reflectivity versus social contracting. *The social institutions of capitalism: evolution and design of social contracts*: 127.
Wilson, D. S. 1975. A theory of group selection. *Proceedings of the National Academy of Sciences of the United States of America* **72** (1): 143.
Wilson, D. S., and E. Sober. 1994. Reintroducing group selection to the human behavioral sciences. *Behavioral and Brain Sciences* **17** (04): 585–608.
Wilson, R. K, and C. M Rhodes. 1997. Leadership and credibility in n-person coordination games. *Journal of Conflict Resolution* **41** (6): 767–791.
Woodward, J. 2003. *Making things happen: A theory of causal explanation*. Oxford University Press, USA.
Wu, Jia-Jia, Bo-Yu Zhang, Zhen-Xing Zhou, Qiao-Qiao He, Xiu-Deng Zheng, Ross Cressman, and Yi Tao. 2009. Costly punishment does not always increase cooperation. *Proceedings of the National Academy of Sciences* **106** (41): 17448–17451.
Xiao, E., and D. Houser. 2005. Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences of the United States of America* **102** (20): 7398.

Zahavi, A. 1975. Mate selection–a selection for a handicap. *Journal of theoretical Biology* **53** (1): 205–214.

Zahavi, A., A. Zahavi, A. Balaban, and M. P Ely. 1999. *The handicap principle: A missing piece of Darwin's puzzle*. Oxford University Press, USA.

Zhu, J. 2004. Intention and volition. *Canadian journal of philosophy* **34** (2): 175–193.

Zimmerman, M. J. 1985. Sharing responsibility. *American Philosophical Quarterly* **22** (2): 115–122.