Writing social reality into the book of the world

Aurélien Darbellay

Writing social reality into the book of the world

Ph.D Thesis

Doctorando: Aurélien Darbellay

Directores: Dan López de Sa Medina & José Antonio Díez Calzada

Tutor: José Antonio Díez Calzada

Programa de doctorado: Filosofía Analítca

Universidad de Barcelona (UB)

Facultad de Filosofía

Acknowledgments

It has been six years since I started this PhD. Over this time, the life of a LOGOS grad student brought me to meet countless people with whom I had both socially and philosophically rewarding exchanges. I'm thankful to all of them for contributing to making these years so exciting and enriching.

I'm grateful to the Gobierno de España for funding my research from 2012 to 2016, through an FPU scholarship. It is beyond reasonable doubt that I couldn't have written this thesis – and benefit as I did from the philosophical training I received – without this economical support.

I'm grateful to Chiara Panizza, Paco Murcia, Pablo Río, Immaculada Murcia Valcarcel and Oscar Cabaco for all the help they've provided me with over the years in solving complex administrative and informatic issues.

I'm grateful to Max Kölbel and Manolo García-Carpintero who managed to secure me funding before I was awarded my scholarship.

I'm grateful to all the people in LOGOS – and specially its core members, whatever that means – for creating a context in which a felt challenged and safe at once. I'll keep remembering the group as a model of organization in which knowledge is shared, students are trained and ideas are criticized without excessive reservation in a surprisingly egalitarian, friendly and horizontal manner.

I'm grateful to the people with whom I shared the sessions of the well-named PMS and SM. Discussions were incredibly stimulating there, and the atmosphere particularly friendly. The feedback I received on the many papers I presented in these seminars have been more than useful. Special mention go to Sven Rosenkranz, Roberto Loss, Marta Campdelacreu (for being the eternal knight of dualism in the middle of the monist ocean), Giuliano Torrengo and Esa Díaz-León.

In the spring 2014, I had the great opportunity to make a stay in Berlin – equally funded by the FPU scheme of the Gobierno de España. I'm grateful to the people who received me there Tobias Rosefeldt,

Alexander Dinges, Julia Zakkou, Daniel James, Miguel Hoeltje, Matthias Raphael, Catherine Diehl. Special mention to Raphael Van Riel, both for all the philosophical discussions and the friendship.

From the very beginning, I had the chance to be surrounded by wonderful fellow grad students, who contributed to make my education more intellectually stimulating and lively. I'll be forgetting most of you but, for what is worth: thanks Carlota, Mar, Romina (you got me the FPU!!!), David, Stephan and Samuele.

I'm grateful to my friends who understood that I needed to disappear during the last months, specially Jessica, Pau, Yoann, Aurélien and my brother Florestan.

I'm grateful to Max Kölbel because, weren't for him I would probably not have jumped into the PhD adventure.

I'm grateful to John Horden for being so smart, so thorough and so obstinate in bringing philosophical conversations to their rightful end. Over the years, you've been a challenge, a model and a constant source of philosophical knowledge.

I'm grateful to José Díez for supporting my FPU application, being there when needed during these years and offering careful and insightful comments and critiques on an earlier draft of this dissertation.

I'm grateful to Ljuba for so many things that I couldn't even dare start drawing a list. Meeting you is definitely one of the best things that doing this PhD brought me.

Je remercie ma famille – Michèle, ma mère, Claude, mon père, et Florestan, mon frère – de m'avoir aidé à rester debout et continuer à travailler quand, deux jours avant la reddition, j'ai eu l'impression que le monde me tombait sur la tête.

I'm super grateful to Dan López de Sa. Mainly for taking, over all these years the task of training me so extremely seriously; for being tough when needed, even though you would have to face my frustration; for being enthusiastic almost all the time; and for giving me incredibly precious pieces of advice. Discussions with you weren't always funny, for I most of the times ended up convinced that I was just

plainly wrong, but they were certainly deeply fun and educative. You strived very hard to make me a better philosopher and, when I compare my writings today with those of three years ago, I'm deeply grateful for the path you helped me walking. Finally, thank you also for the support you gave me during the weeks previous to my deposition. You took me out of a cave more than once.

Finalmente, te agradezco María, tu paciencia, tu amor y tu atención durante estos últimos meses en los que te ha tocado vivir con un zombi que trabajaba todo el tiempo. Has sabido estar cuando te necesitaba; y aceptado no estar cuando te lo pedía. Y has tolerado – sin quejarte demasiado – mi preocupación casi constante y mis recurrentes ausencias. Has encontrado las palabras justas, los consejos adecuados, y te has mantenido firme cuando hacía falta. Gracias por todo eso.

Contents

Summary

This thesis has four chapters which, together, offer a non-exhaustive but still quite extensive account of social reality.

In Chapter 1, I defend an analysis of the notion of sociality which has it that to be social is to somehow involves cooperation.

In Chapter 2, I present and defend an account of cooperation which roughly claims that cooperation occurs when people pursue a goal they have in common, according to a plan they share, and in a state of mutual awareness.

In Chapter 3, I put forward an account of a particular kind of social phenomena: institutions. I draw on the work of Searle, Hindriks and Thomasson and argue, very roughly, that institutions are created by people who cooperate in a way that commits them to certain rules, which they have the power to enforce.

Finally, in Chapter 4, I argue that circular accounts of social phenomena can be illuminating and, in particular, that they can help vindicating reductive identifications of the *phenomena* they target, even though they fail to provide reductive identifications of the *notions* they aim to analyze.

Introduction

Writing social reality into the book of the world[1] is, as I understand it, offering a description of social reality which makes explicit its relations to other, plausibly more fundamental, aspects of reality. Accordingly, the aim of this thesis is to offer a (non-exhaustive but yet extensive) description that satisfies this desideratum.

On this reading, the project of writing social reality into the book of the world is by no means original. It is, for instance, a project in which both John Searle (see (Searle, 1995) & (Searle, 2009)) and Michael Bratman (see (Bratman, 2009) & (Bratman, 2014b)) have explicitly invested much effort. And, to a lesser extent, it is also a project that can be attributed to Margaret Gilbert (see (Gilbert, 1992)). My contribution to this endeavor heavily draws on the work of these predecessors (specially Bratman and Searle). But it also differs from their proposal in significant respects. From Searle, I inherit the idea that the hallmark of social reality is cooperation. But, unlike him, I do not claim that cooperation involves a special form of collective intentionality, irreducible to individual intentionality. Like Bratman, I try to build cooperation out of individual actions and intentions. But, as it will be argued in due time, the elements I use to this effect differ in interesting respects from the one he appeals to. Finally, while I endorse Searle's view that institutions are systems of rules, I offer a radically different account of the way how people create the institutional rules they have to live by.

My plan is as follows. In Chapter 1, I address the first question that one needs addressing in order to locate social reality in a broader world-picture: what is social reality? I answer this question by putting forward a conceptual analysis of the notion of sociality according to which to be social is to involve (in a sense that I discuss at length) cooperation.

[1] The title was inspired by Sider's *Writing the Book of the World* (Sider, 2014). I should flag here that I do not intend to imply that I endorse the view Sider defends in this book.

Given the conclusion of the first chapter, the next step towards a description of social reality that would make explicit its relations to other aspects of reality is to build an account of cooperation. I devote Chapter 2 to this endeavor. In a nutshell, the account reads that cooperation occurs when people pursue a goal they have in common, by following a plan they share, and in a state of mutual awareness. The notions of common goal and shared plan are carefully defined to make explicit how they are constructed out of the intentions of individual agents. Hence, the account shows how basic social phenomena (i.e. cooperative activities) can emerge out of other, plausibly more fundamental, aspects of reality.

In Chapter 3, I extend my description of social reality by offering an account of institutional reality. Roughly stated, the account has it that people create institutions by cooperating in a way that commits them to certain rules, which they have the power to enforce. I argue that the account is reductive in the sense relevant to the overall purpose of this thesis.

Chapter 4, finally, is better seen as a methodological contribution. It consists in a discussion of circular accounts of collective phenomena and suggests that, their limitations notwithstanding, such accounts can be insightful in many respects. My motivation for this discussion is double. On the one hand, I take it that, in spite of appearances and despite my efforts to avoid it, there is room to argue that the accounts of Chapter 2 and 3 are circular (or, in the case of the account of Chapter 2, that we have to interpret it in a way that makes it circular in order to avoid counterexamples). On the other hand, it is sometimes assumed in the literature that circular accounts are unfit to back up the reductive identification of collective actions and intentions to sums of individual actions and intentions. But this claim, whose bearing on the overall aim of my thesis is – I believe – clear enough, is mistaken: even though circular account fails to provide reductive identifications of the *notions* they target, they can offer reductive identifications of the *phenomena* these notions refer to. Or so I argue.

Given the nature of the project, shortly stating the conclusions of my thesis is uneasy. But, as matter of synthesis, I shall offer the following: cooperation, a phenomenon which consists of interrelated

individual actions and intentions, is the social atom; one way or another, it is involved in the emergence of every social phenomena; in particular, when suitably combined with power, it gives rise to institutions.

Chapter 1: Sociality[2] as cooperation

Abstract: In this Chapter, I present an analysis of the notion of sociality which I label SOCIALITY AS COOPERATION. According to this account, the social is aptly characterized as that which somehow involves cooperation. I start by motivating the general thesis, as unspecific as it is. Then I tackle the question of how we should interpret the phrase "somehow involve". I argue that the notion which best serves our purposes is one construed in terms of essence.

1.1. Introduction: "social", "social reality", etc.

In this chapter, I aim at offering an account of what is the social, or – alternatively – an analysis of the notion of sociality. Now, as we take up this project, a few comments are needed to pin down our target: for the expression "social" and its likes are used with many different purposes. Thus, for instance (and without trying to be exhaustive):

(1) "social" is sometimes used to mean that a feature of a human population is *contingent* upon human's biological nature, i.e. that humans could fail to exhibit such a feature. This seems to be (at least) one of the senses relevant when people claim that gender is social.

(2) in "social policies" and "social work", "social" appears to be used to signify that something aims at human welfare.

(3) "social" is sometimes used as a synonymous for sociable: thus, someone might be said to be very social if she enjoys being surrounded by other people, etc. (see (Gilbert, 1992)).

(4) "social" is sometimes used to say that a phenomenon somehow involves agents interacting with each other. This seems to be the sense at work, when we say that language, money, borders, leaders and hierarchies are social phenomena.

[2] « Socialness » or « sociality »: both seem acceptable to me. I chose the second one in deference to Gilbert's pioneer work on the topic (e.g. (Gilbert, 1998)).

It is this last notion of sociality I'm interested in. This notion grounds a two-fold partition of reality. That is, in this sense, everything either belongs or fails to belong to social reality. Furthermore, according to this notion, the destruction of the Roman Empire, the start of the Industrial Revolution, the average Salary in Spain, the fact that Spain is a Kingdom, and the fact that Bill Gates is richer than me are arguably *social* facts, phenomena or objects. On the other hand, the extinction of dinosaurs, the beginning of the Huronian Ice Age, Mt. Everest, the average life span of a star, the fact that genes are made of DNA and the fact that grass is green presumably do not belong to social reality.

It is an assumption of mines that, in the lines above, I'm not coining a new notion, but rather gesturing towards a concept of sociality that my reader, as a member of the English-speaking community, will likely be acquainted with. In this chapter, it shall be my contention that to be social (in this sense) is to involve cooperation (in a way that shall be made more precise later on).

1.2. The nature of the social[3]

I claimed above that to say of something that it is social – in one of his ordinary usages– is to say that it somehow involves agents interacting with each other.[4] I think that this is roughly correct. But it is also incomplete. On the one hand, it is nothing but fair to ask for an (at least partial) elucidation of that which "somehow involves" stands for (and this I shall do in Section 1.3). On the other, it is pretty clear that not any old interactions are social. Thus, for example, when a lion hunts down a gazelle, it is uncontroversial that an interaction takes place. But it isn't a social phenomenon. Another example:

[3] Surprisingly enough, despite the growing interest in social ontology over the past years, the question of what makes certain things *social* hasn't received much attention (for a similar diagnosis, see (Mason, 2016, pp. 841–842)). In his seminal work, Searle defines "social facts" as those facts "involving collective intentionality" (see (Searle, 1995, p. 26)). But he regards this definition as a stipulation and therefore does not bother to defend it. As for Epstein, although he grants that "To be sure, a distinction should be made between the properties that count as "social" and those that do not", he doesn't offer an account of this distinction and simply states that "it seems likely that [the] standards are low – all that is needed is a little social salt added to the generative stew" (see (Epstein, 2014, p. 68)). Finally, Margaret Gilbert has defended throughout her work the view that the hallmark of the social is a phenomenon she labels "joint commitment" (see in particular (Gilbert, 2003) and (Gilbert, 2013)). Insofar as, in her view, joint commitment is crucially involved in cooperation, the view she defends may not differ that much from SOCIALITY AS COOPERATION as I defend it in this chapter.

[4] I work with an undemanding notion of agent according to which anything capable of performing actions is an agent.

two tigers, Herb and Oscar, want to eat the same dead pony; they fight; Oscar wins and happily calms his hunger. Herb and Oscar interacted and there was nothing social about it. And even if instead of two we picture a hundred of tigers fighting over the corpse of a pony, we still fail to see any social going-on.

Thus, not all interactions are social, not all intraspecific interactions are social, and not all intraspecific interactions involving many agents are social. Rather, I submit, interactions are social phenomena when they are instances of *cooperation*. There is nothing social about a lion hunting down a gazelle, for their interaction isn't cooperative at all: rather, they pursue incompatible goals and each tries to prevent the other from achieving its. And the same is true of Herb's and Oscar's interaction. On the other hand, as soon as cooperation raises its nose – say two lionesses help each other to capture a buffalo – some kind of social phenomenon seems place. Motivated by these preliminary observations, I suggest that we consider the following proposal:

SOCIALITY AS COOPERATION: to be social is to somehow involve cooperation

No doubt, SOCIALITY AS COOPERATION is, as it stands, *prima facie* quite appealing. Cooperation is pervasive across human societies. We cooperate when we elect a president, when we build dams and buildings, when we dance, when we communicate, when we play soccer against each other, or even when we box, etc. Furthermore, for many (or most, or all) paradigmatic social kinds, it is uncontroversial that they somehow depend upon cooperation. How could there be judges, if people didn't cooperate in maintaining a legal system; how could there be money, unless a given system of exchange was kept functioning by agents cooperating to this end, etc. Finally, we describe as social just those animals which systematically cooperate in breeding, dwelling, hunting, etc.

Nevertheless, SOCIALITY AS COOPERATION faces several challenges. Hence, for instance, some presumably social facts (e.g. that women are more likely than men to wear pink clothes) do not obviously involve cooperation; and the same is true of some social kinds (e.g. war). Furthermore, some may worry that SOCIALITY AS COOPERATION is hard to wed with the role of conflict in shaping

societies. I shall try to dispel these worries. But before doing that, and in order to avoid misunderstandings, I need to spend some time discussing the key notion of cooperation.

1.2.1. Cooperation

First a disclaimer: do not expect to find an account of cooperation – in the sense of an analysis thereof – in the following paragraphs. I develop such an account in Chapter 2. Rather, in the present chapter, I wish to defend SOCIALITY AS COOPERATION independently of any particular elucidation of the nature of cooperative phenomena. I believe this is the right way to go, since any such elucidation shall be controversial. Thus, the aim of the following lines is merely to make sure that we have the same phenomenon in mind, when we speak of cooperation.

Paradigmatic examples of cooperation are found in actions performed together by several human beings (see (Searle, 1990), (Gilbert, 1990), (Tuomela, 2000), (M. E. Bratman, 1993), (Kutz, 2000), (Ludwig, 2007), among others). Thus, my brother and I cooperate with each other every time we go for a walk together, prepare a mayonnaise together, clean a park together, and paint a house together, and so on....

Cooperation also happens at a bigger scale: millions of people cooperate (or should cooperate) in protecting the environment by systematically separating waste, just as millions of people cooperate by following the Highway Code of the country they live in, and millions of people cooperated across the world in the massive protests against war on Iraq in 2003. Furthermore, it would seem that social institutions (e.g. monetary and judiciary systems) require for their continuous existence the cooperation of most members of the society they help structuring.

Albeit cooperation always involves several inter-related agents, not any interaction is an instance of cooperation (e.g. a lion hunting a gazelle), not even among humans (e.g. many cases of street fights – those who aren't held according to any rule). In particular, cooperation should be differentiated from (mere) strategic interaction, where each participant chooses a given course of action partly on the basis of her expectations regarding the courses of action that the other participants will pursue. It is

an open question whether cooperation always involves strategic interaction, but there are certainly non-cooperative instances of the latter (e.g. John wants to avoid Jack, who wants to see John: they strategically interact, since each of them makes decisions on the basis of what he expects the other to do. But they don't cooperate).

As suggested above, cooperation is also closely related to the phenomenon of doing things together. Many instances of doing things together are instances of cooperation, e.g. going for a walk together, playing tennis together, etc. Furthermore, cooperating is a paradigm of something that agents do together. But these may nevertheless be two different phenomena: arguably, agents can do something together without cooperating in doing it. For instance, it may be claimed that humans cause global warming *together* (after all, this is definitely something they do – and yet, none of them seems to be doing it on its own); and yet they clearly do not cooperate in doing so. On a different line, one may argue that the following scenario also contains an instance of non-cooperative doing together. I'm the boss of a company. My employees don't know what the company's goals are. They are furthermore not allowed to directly interact with each other. They receive instructions which they have to blindly follow, if they want to get paid. Whatever the company does, we arguably do it together; and yet, it would be strange to describe us as cooperating in doing what the company does.[5]

Presumably, cooperation also exists in the non-human reign: many mammals (hyenas, wolfs, lionesses etc.) appear to hunt in a cooperative fashion, for instance. More controversial is the question as to whether so-called social insects offer genuine (as opposed to metaphorical) examples of cooperation. The idioms of cooperation come readily to the mouth of those who describe the life of, say, ants (for instance, it is common to read that ants practice cooperative breeding (Crespi & Yanega, 1995)). But some may be tempted to argue that, in such contexts, either the notion is used metaphorically, or it is a different – albeit related – notion that is being appealed to. Deciding on the issue arguably requires,

[5] (Chant, 2006) argues that there are instances of unintended collective actions. If she is right, most (if not all) such instances would be non-cooperative collective actions.

on the one hand, a conceptual analysis of cooperation and, on the other, empirical work to determine whether so-called social insects meet the conditions revealed by the conceptual analysis.

1.2.2. Sociality as cooperation: an unsystematic defence

It is now time to defend and argue for SOCIALITY AS COOPERATION. I should flag from the very beginning that I do not have a systematic argument in favor of the thesis. Rather, I came to be convinced by going through many examples and observing, in each case, how well the thesis accommodates them. In what follows, I try to convince my reader in the same, unsystematic and piecemeal way.

I shall start by focusing on the non-social – which, according to SOCIALITY AS COOPERATION, shouldn't involve cooperation. Clearly enough, paradigmatic examples of entities which do not belong to social reality do not involve cooperation either: thus are mountains, rivers, trees, stars, supernova, solar systems, and the beginning of the Huronian Ice Age. This doesn't bring much support to the thesis though: for these entities do not only fail to involve cooperation, they do not involve agents at all; and this seems to be enough to account for their non-sociality.

Slightly more interestingly, we may observe that lonely agents – agents who do not belong to any kind of society or social groups – are also agents who fail to cooperate. In particular, the Crusoe-like fictions we imagine when we want to picture what kind of a man could develop without any social surroundings are invariably cooperation-free scenarios. Prima facie, the support that SOCIALITY AS COOPERATION receives from such observations is once again quite meager: for, apparently, these fictions only involve *one agent* (as opposed to several); and this fact alone would appear to explain satisfactorily why they do not contain any trace of social reality. But that's not quite right: these scenarios may perfectly involve more than one agent; the lonely human being might be pictured as surrounded by wild dogs, apes, birds, etc. without taking part in any kind of social phenomena (that is without being less *lonely* in the sense pointed out above). Thus, if Crusoe and her likes lack social surroundings it isn't because they are *the only agent in the vicinity*.

Now, it may be suggested that the reason why Crusoes[6] don't partake in any kind of social phenomena is that they have no relation to agents *of the same species*, thus implying that social phenomena must be somehow homogeneous with respect to species. There is an obvious problem with this suggestion, though: societies and social phenomena in general do not have to be species-homogeneous. It doesn't take much imagination to picture societies composed by members of different species (science-fiction offers countless examples thereof). And, indeed, Moogly-like (or Tarzan-like) versions of the story of the wild man precisely picture human beings engaging in social relations with agents of different species (wolfs in the case of Moogly, apes in the case of Tarzan). Hence, the fact that Crusoes aren't related to agents of the same species doesn't suffice to explain why they do not participate in any social phenomenon.

It is worth emphasizing furthermore that Crusoes may have quite rich relations with the wildlife around them without hereby giving rise to social goings-on. Thus, Anabelle (Crusoe) and Robert (the Chimp) may both be aware of their respective presence on Zanator (an Island in the middle of the see); they may furthermore frequently interact (if, for instance, Anabelle wants to eat Robert, and Robert is afraid of Annabelle, and they monitor each other's behavior and choose what to do depending on what they expect the other to do); and they may do all of this without even getting close to make it the case that there is social life on Zanator. Hence, if Crusoes don't belong to any kind of society, it isn't because they do not interact with other agents – or because they do not interact with them frequently enough.

Finally, even adding other human beings to the picture doesn't automatically secure the presence of social phenomenon. It doesn't make much of a difference if, instead of Robert the Chimp, it is Alfred the human being, who shares Zanator with Annabelle: as long as their interactions are purely antagonistic (Annabelle tries to eat Alfred, and Alfred tries not to be eaten), Zanator shall remain a social desert. On the contrary, as soon as Annabelle and Alfred come to terms (say Annabelle decides that Alfred is more valuable as an ally than he is as a dish, and Alfred realizes that) and start doing

[6] I'll use the plural "Crusoes" to refer to Crusoe and her likes.

things together (e.g. hunting, building shelters, looking for a way out of Zanator, etc.) their behavior and their relation become properly social. At the end of the day, then, there are reasons to believe that SOCIALITY AS COOPERATION is the only (natural and initially plausible) thesis capable of explaining why Crusoe and her likes lack any kind of social surroundings. And this, I daresay, brings more than a meager support to the proposal.

After this excursus on the edge of social reality, I shall now steer towards its core, to consider and discuss several paradigms of social kinds. For each of them, I'll argue that they involve cooperation just like predicted by SOCIALITY AS COOPERATION.

> *Money*: I take it to be uncontroversial that the nature of money is, at least partly, to be a medium of exchange. And, again uncontroversially, exchanging is a cooperative activity. Thus, uncontroversially, money somehow involves cooperation.
>
> *Inflation:* inflation is, by definition, a sustained and rapid increase in prices, and a correlated devaluation of a certain currency. But currencies are money and inflation therefore somehow involves money. And hence, since money involves cooperation, so does inflation.
>
> *Leader*: this example of social kind has been used to try and undermine the claim that we-attitudes (the attitudes characteristically involved in cooperation) are needed to underpin institutions (see (Ylikoski & Mäkelä, 2002)). And some may think that it also causes trouble for SOCIALITY AS COOPERATION. For, trivially, the nature of leadership is to lead a group of people and, at least *prima facie*, it would seem that leading isn't a cooperative activity. Thus it isn't at all clear how (and whether) cooperation is involved in leadership. And then SOCIALITY AS COOPERATION is in trouble. This argument fails, though; for it is only when we look at the matter superficially that we can be driven to believe that leading isn't itself a cooperative activity. The leader and the led cooperate in an action whose nature is just this: to have a leading and a led part. Intuitive support for

this claim comes from considering cases in which someone gets other people to do what she wants and yet, we would refrain from calling her a leader. Say for instance that I know very well what Lara, Sarah and Jerôme want. Based on this knowledge, on their relative simple-mindedness and my conviction power, I very often manage to have them doing what I want: I convince them that the best way to serve their purposes is to perform actions of which I know – without telling them – that they will serve mine. In this scenario, I have Lara, Sarah and Jérôme doing what I want and, yet, I'm no leader. My suggestion is that, among other things, what is lacking here is that the way in which I influence their decisions doesn't pertain to a scheme that we (cooperatively) maintain together. When I dishonestly convince Jérôme to do something that is good for me, I do not feel in any sense that my behavior relates to or depends upon a collective endeavor: it is only part of my plan to get the most out of my options. Same goes for Lara, Sarah and Jérôme: they follow my 'advices' out of completely individual considerations. But, on the other hand, I would argue that, if I'm a leader, people understand that, when they do what I tell them to, they are cooperating with a group who follows the same rule. In other words: they obey because "that's what we do".

War: some will be tempted to consider war as an obvious counterexample to SOCIALITY AS COOPERATION. For the nature of war seems to crucially involve both fighting and antagonistic interests, two elements whose presence appear to be incompatible with cooperation. But on a closer look, we quickly realize how mistaken are these considerations. For, war doesn't oppose lonely individuals – not even a million of people can fight a war, if each of them is fighting on her own. Rather, war oppose groups of people (countries, nations, religious communities, and what-have-you); that is, in war, each side is constituted by many people who *cooperate* with one another.

Competitive games: some may think that competitive games constitute another counterexample to SOCIALITY AS COOPERATION. But this is once again mistaken: for, suppose María and I are playing chess, then no matter how hard I try to beat her, I must be cooperating with her in following the chess's rules if we are to be playing chess at all.

Institutions: Institutions involved cooperation in many ways. Hence, as Searle has argued in several places (see (Searle, 1995, p. 39), (Searle, 2009, p. 58)) that many transactions which depends on the existence of institutional structures (buying something, communicating with the help of a conventional language, showing one's ID to get authorized to cross a border, signing a contract, getting married, playing a game, etc.) are essentially cooperative. In other terms, he claims that many of the functions (being money, being a judge, being a border) institutions create can be "performed only as a matter of human cooperation" (Searle, 1995, p. 39).[7] But this doesn't suffice to show that institutions in general involve cooperation – I believe. For there are arguably institutional functions whose performance doesn't require cooperation. Consider, for instance, the institutional status "being a restricted area". Say, for instance, that it is a rule of Zambesi (a small village) that the forest laying Southwest to the village is a restricted area: none is allowed to enter it. As I see it, it is less than obvious that this status has a function which can only be performed as a matter of cooperation. Hence, it would seem that if each inhabitant of the village systematically refrains from entering the forest, then whatever function might be attached to the status "being a restricted area" is effective or performed – and it isn't clear how does this performance requires cooperation. More fundamentally and generally, though, I take it that institutions involve cooperation because they are created by commitments people acquire by engaging in cooperative activities (I shall argue for this claim at some length in Chapter 4).

[7] Indeed, his claim is that all the functions created by institutions require a cooperative effort for their performance.

It is my contention that, once we look deep enough, we will find that every single bit of social reality somehow involves cooperation pretty much like money, inflation, competitive games, institutions, leadership and war do. I could support this claim by going through many more cases. But I take it that such an exercise would be of little value: I have at any rate no space to cover sufficient ground and prove that no counterexample could be found; and the examples considered so far suffice, I believe, to make SOCIALITY AS COOPERATION very plausible.

To conclude my unsystematic defense, I shall now address two objections that are likely to be raised against the proposal I put forward here.

According to a first objection, SOCIALITY AS COOPERATION offers an overintellectualized picture of social reality. For, my objector has it, cooperation is a very demanding phenomenon – one which requires a relatively sophisticated theory of mind. But, the objection continues, it is fairly implausible that the animals which we describe as social have such a sophisticated theory of mind. Hence, a weaker notion – such as maybe coordination – would probably constitute a better candidate.[8] I'll discuss at some length the claim that cooperation requires a sophisticated theory of mind in Chapter 2. But, for the time being, I don't need to challenge this claim to answer the objection. For, as a matter of fact, we naturally describe as engaging in cooperative activities those animals which we reckon as social. Hence, as already put forward in Section 1.2, it is nothing but natural to say that two lionesses cooperated to hunt down a gazelle, or that some wolves cooperated to protect their offspring from a predator. Indeed, we even use the idioms of cooperation to describe the life of social insects (see (Crespi & Yanega, 1995)). And this suggests that the conceptual connection that SOCIALITY AS COOPERATION highlights isn't threatened by considerations to the effect that some animals, that may not have a sophisticated theory of mind, nevertheless engage in social phenomena. Granted: it may be that, *strictly speaking*, cooperation requires a sophisticated theory of mind. And, accordingly, it may be that, *strictly speaking*, the claim that wolves, lionesses and ants cooperate is false. But then, I take

[8] This objection was raised by Sally Haslanger.

it that it would be plausible to say that, *strictly speaking*, wolves, lionesses and ants do not engage in social phenomena either.

Furthermore, coordination is, I take it, unfit to characterize the social. Consider: the As and the Bs are two species of lonely animals (they don't live in groups) which migrate twice a year. The As mainly eat the Bs' feces. As for the Bs, they mainly eat a type of fruit which only grows in presence of the As. As a result of this alimentary interdependence, the As' and the Bs' migration are *coordinated*: both the species are hardwired to start migrating in response to the same clues, which ensures that they will find food on their way. And yet, there seems to be nothing social going on here.

As for the second objection, it complains that SOCIALITY AS COOPERATION conceals the pervasiveness and importance of conflict in the shaping of societies, their structures and institutions. I guess that this worry might be particularly present among those who have sympathies for the so-called *theories of conflict* in sociology (a tradition fathered by Karl Marx).

Now, it is not obvious to me why the thought that conflict is basic in shaping our societies should be hard to wed with SOCIALITY AS COOPERATION. My best guess is that the argument would go more or less as follows. Firstly: conflict requires antagonistic interests. Secondly: cooperation requires common interest. Thirdly: if the basic structures of societies are kept functioning in a cooperative fashion, as SOCIALITY AS COOPERATION implies, the basic structure of societies are shaped by common interest, rather than conflict. Finally: conflict is the force that shapes societies, so SOCIALITY AS COOPERATION is wrong.

To this objection from conflict, I offer the following reply: the kind of common interest that cooperation requires – if any – is such that the agents who cooperate with each other can perfectly be in conflict with one another. Even more, the object of their conflict can be the very cooperative activity that they engage in together. Consider the following story: Hillary and Maude were buddies. They used to rob banks together. Hillary wants to rob a last bank. She wants Maude to jump in. Maude doesn't want to. Hillary insists more and more, and the conflict becomes violent. At a certain point, Hillary

kidnap's Maude's wife and threaten to kill her if Maude doesn't participate in a last bank robbery. Eventually, Maude accepts and they proceed to rob a last bank together. Let us assume for the sake of the argument that, as they work together on the robbery (i.e. as they *cooperate* in preparing and carrying out the deed), Maude and Hillary indeed have a common interest. It is clear that such a common interest doesn't make the conflict that tears them apart from each other less dramatic in any sense.[9,10]

This being so, *conflict theories* need not reject SOCIALITY AS COOPERATION. Yes, my proposal says that social structures and institutions are kept alive by people who cooperate with each other. But this is perfectly compatible with the presence of very harsh conflicts regarding which are the cooperative activities that should be undertaken (and how they should be undertaken, and whether there should be cooperation at all between the people in question).

1.3. "somehow involving"

According to SOCIALITY AS COOPERATION, to be social is to *somehow involves* cooperation. But how should we understand this expression "somehow involve"? In the following, I'll consider three different options: a causal, a modal (based on either necessitation or supervenience) and an essential understanding (or specification). I'll argue that it is the essential reading that we should build into our account.

First thing first, though, I need now to say a bit more about the way how I understand SOCIALITY AS COOPERATION. The thesis reads:

[9] In his *Shared Cooperative Activity*, Bratman claims that cooperative activity (which, I assume, is identical to cooperation) is incompatible with the kind of coercion involved in this example (see (M. E. Bratman, 1992, pp. 334–335). His argument to this conclusion is that, when she coerces Maude into forming the intention that they rob a last bank together, Hillary's attitude isn't cooperative at all. I agree with this much: Hillary's coercing Maude isn't a cooperative activity. But I fail to see why this should entail that Hillary and Maude's robbery isn't a cooperative activity. After all, the coercion and the robbery are different activities, even though the latter is a result of the former.

[10] The point that cooperation is compatible with coercion has been made by Tuomela (see (Tuomela, 1993, p. 99).

SOCIALITY AS COOPERATION: to be social is to somehow involve cooperation.

But, one may wonder, which is the intended strength of the identity the thesis expresses? Is it a mere extensional equivalence (that would be satisfied if, in the actual world, everything social somehow involves cooperation)? Is it a metaphysical equivalence (that would require that, in every metaphysically possible world everything social somehow involves cooperation)? Or is it something else?

I intend SOCIALITY AS COOPERATION as providing a conceptual analysis of the notion of sociality (or being social). As such, I take it that it should be read as a *conceptual* identity or equivalence – that is, along the lines of:

- The concept of being social is the concept of somehow involving cooperation
- To say that something is social is to say that it somehow involves cooperation
- To think of something as social is to think of it as somehow involving cooperation
- social $=_{def}$ somehow involves cooperation

I have no particular view regarding the way how such conceptual equivalences should be stated. In what follows, I'll use the last version – but only because I like the aspect of it. Hence, anyone should feel free to read SOCIALITY AS COOPERATION as stated how she thinks that conceptual equivalences are best stated. Hence, we have:

SOCIALITY AS COOPERATION: social $=_{def}$ somehow involves cooperation

The aim I pursue in this last section is to find the interpretation of "somehow involves" that makes SOCIALITY AS COOPERATION so understood true.

1.3.1. The causal understanding

Quite uncontroversially, many social facts and events (if not almost all of them) causally depend upon cooperation. For instance, in the causal history of any strike, any election, and many social practices (such as driving on the left/right, using certain items as medium of exchange, etc.), there are people

engaging in conversations. And this kind of communication among humans (conversation) is a paradigm of cooperative phenomenon. Generalizing on this initial datum, we might want to assess the merits of a causal understanding (or specification) of SOCIALITY AS COOPERATION:

Causation: social $=_{def}$ depends causally upon cooperation.[11,12]

I find **Causation** intuitively unappealing. Think of money, for instance. Is it the case that, when we say that money is social we merely mean that money causally depends upon cooperation? Isn't there something more to money's sociality? Something along line of the idea that money's existence is *constituted* (or grounded) by people cooperating to perform their transactions according to certain rules? Or think of language. Does the claim that language causally depends upon cooperation appropriately capture the thought that language is social? Doesn't it rather embody the idea that languages are, by definition (or nature), things that people use to communicate (a cooperative activity if any)? And when we say that cooperation is a social phenomenon, do we really mean that cooperation depends causally upon cooperation? Finally, do we really want to say that any old consequence of cooperative activities is social as **Causation** appears to imply (more on this below)?

In order to go beyond these preliminary remarks, let me highlight two obvious consequences of **Causation**:

(1) Everything that causally depends upon cooperation is social.

(2) Everything social causally depends upon cooperation.

I think that both (1) and (2) are faulty. In what follows, I argue first against the later and then against the former.

[11] Where X causally depends upon Y if and only if Y is among X's causes.

[12] Obviously enough, a final version of **Causation** would require some finessing (e.g. it would say that social kinds causally depend upon cooperation *for their instantiation*, etc.). But the details of the refinement are irrelevant to my discussion, and I thus set them aside.

My argument against (2) is simple: when paired with an uncontroversial principle about causation and the platitude that cooperation is a social phenomenon, it generates a regress with unacceptable consequences. The principle in question states that, if Y partly causes X, then Y is temporarily prior to X. From there, it doesn't take much imagination to come up with the regress: take any social event, call it O; if every social event causally depends upon cooperation, so does O; thus, by the principle, an episode of cooperation E is temporarily prior to O; but any episode of cooperation is a social event; thus, if every social event causally depends upon cooperation, there is an episode of cooperation E' prior to E; and so on…

Given the regress, anyone who wishes to endorse (2) must adopt one of the four following claims: (i) always some social event occurred in the past, (ii) there is no social event, or (iii) cooperation isn't a social phenomenon, (iv) the uncontroversial principle about causation is false. Going for any of the three first options is definitely not an option (at least in the present context, with respect to (ii) – in other contexts, eliminativism about social events may be an option, although not a very appealing one, I believe). As for the third, I would argue that none should accept such a claim in order to stick to her favorite account of sociality. Thus, (2) should be rejected.

Some may feel tempted to object that,[13] on the plausible enough assumption that time is dense in itself (i.e. for any two points in time, there is another point in time between them), the regress is unproblematic. For if time is dense in itself (and on the assumption that something can be cooperative no matter how short-lived it is[14]), then – the objection goes – each cooperative going-on that happens after t_0 could be caused by an "older" cooperative going-on which *also happens after* t_0.[15] Thus the

[13] This objection was brought to my attention by Roberto Loss.

[14] The objection could, I believe, already be resisted at this stage. But I'll grant the assumption to my opponent and focus on another reply.

[15] Let us start with cooperative going-on C_1, located at t_1. Let us then name "t_2" the time that is exactly in the middle between t_1 and t_0. Assume there is a cooperative going-on located at t_2, and call it C_2. C_2 could be a cooperative causal antecedent of C_1, thus satisfying (2). Then name "t_3" the time that is exactly in the middle between t_2 and t_0. Assume there is a cooperative going-on located at t_3, and call it C_3. C_3 could be a cooperative causal antecedent of C_2, thus satisfying (2). Repeating the operation over and over again, it seems that we can build a model in which every cooperative goings-on causally depend upon cooperation, and yet cooperation doesn't happen before t_0.

conjunction of (2), the uncontroversial principal about causation and the claim that there is (at least) a social event doesn't imply that always some social event occurred in the past: the whole of social reality may still be located after t_0.[16] Its ingenuity notwithstanding, this objection is mistaken. According to the picture that my objector is drawing, there is a time t_0 in the immediate future of which (no matter how short we take this immediate future to be) there are infinitely many almost instantaneous and causally connected cooperative goings-on. But now, surely, these cooperative goings-on must constitute a single cooperative event. For cooperation takes time; and if there are almost instantaneous instances thereof, they must be the temporal parts of a longer-lived cooperative episode. Let us then call E_0 this bigger instance of cooperation. Even though it doesn't occur at t_0, E_0 is such that for any time after t_0, E_0 starts before that time (for E_0 is constituted by the members of a set which converges towards t_0). Thus, according to the uncontroversial principle about causation, its causes must be located at t_0 or before. Finally, then (2) requires that there be some cooperation going on at t_0 or before.

Others may feel that, once we take vagueness into account, the argument loses its strength. They shall argue that I'm just gesturing at yet another soritical series: as we look back into the causal history of any social event, we will find first clear cases of cooperation, then borderline cases thereof, and, finally, instances of phenomena which are determinately not cooperation. Thus, the objection goes, (2) is all right; or, at least, on as firm a standing as the claim that every human being is born from a human being, or the claim that every bald woman would still be bald if she had one more hair. The obvious problem with this objection is that the last two claims are not on a good standing, quite the opposite: on the most influential strategies to solve the sorites paradox (epistemicism and supervaluationism)[17] these claims turn out to be false. Thus, presenting my argument under a soritical guise doesn't make it any weaker.

[16] Crucially, on this picture, there is no social event occurring at t_0.

[17] For a presentation and defense of the epistemic solution to the sorites paradox, see (Williamson, 1994 CH.7). For a discussion of the supervaluationist solution, see (Dummett, 1975).

I therefore conclude that (2) is false – and therefore so must be **Causation**. Does this suffice to completely exclude the causal understanding of SOCIALITY AS COOPERATION? Not quite, for someone may grant that a causal understanding of "somehow involve" cannot exhaust the content of "being social", and yet insist that sociality is a disjunctive concept. According to this perspective, while some things are social in virtue of involving cooperation in a different – yet to be made explicit – sense, some things are social in virtue of causally depending upon cooperation. Hence, the defender of such a perspective has it that the right understanding of SOCIALITY AS COOPERATION runs along the following lines:

> DISJUNCTIVE SOCIALITY AS COOPERATION: social $=_{def}$ causally depends upon cooperation or involves cooperation in another (yet to be determined) sense.

Such a view still has it that – albeit it isn't a necessary condition – causal dependence upon cooperation suffices to make something social. Hence, it is wed to (1) – the other consequence of **Causation** highlighted above – to which I now turn my attention.

The literature on social constructionism (see (Diaz-Leon, 2013), (Haslanger, 1995) and (Hacking, 1999)) has repeatedly pointed at a sense in which some things are socially constructed in virtue of their causal history. Consider the cases of the Dothrakis: on the basis of a governmental decision, blue-eyed Dothrakis receive a particular education, a result of which being that they are afraid of water. In such circumstances, it seems all right to claim that the fact that blue-eyed Dothrakis are afraid of water is socially constructed. But, beside its causal history, there is nothing to this fact that could ground its sociality: water is a natural kind, fear is a good candidate to be a natural emotion, and being blue-eyed is as natural a feature as people can have. For a closer and more controversial example, consider the claim that women receive a distinctive education, which brings it about that they are less likely than men to enjoy eating raw meat. Anyone accepting such a claim – independently of whether she takes gender categories themselves to be *social* categories – would appear committed to the thought that women's relative non-proclivity towards raw meat is socially constructed.

These observations may be taken to give reason to accept (1), i.e. the claim that causal dependence upon cooperation suffices to make for sociality.[18] However, further reflection reveals serious difficulties with this last claim.[19] Consider: there is a war on the French coast; as they explode, bombs create a violent wind blowing towards the inner Mediterranean Sea; on a nearby island, trees are hit by this wind, which makes their leaves fall to the ground. Uncontroversially, social factors (the war raging on the French coast) play a significant role in causing the leaves' falling; and yet, the fact that, on the island, trees are naked hardly appears to be social in any interesting sense. Or, for another example: some people, the Yuppies, moves to Yande, a location that had remained wild until then; they work together pretty hard to build a new village; the noise they make scares most of the animals around and, when they are done, there is almost no wildlife in the vicinity of the freshly build village. Clearly enough, this latter fact causally depends upon cooperation; and yet, equally clearly, it isn't properly reckoned as a social fact. Overall then, it seems uncontroversial that, as a general principle, (1) should be rejected: there are things caused by cooperation (or other social factors) which aren't themselves social.

Still, one could try to hold on to a restricted version of (1) – one that would avoid obvious counterexamples like the above and still be fit to accommodate the plausible cases of causal social construction. That is, one may insist that causal dependence upon social factors *does* make for sociality, provided that it unfolds in the *right* way. The problem with this strategy is that the most natural candidates to cash out what this right way could be do not do. Requiring that the social factors be non-redundant is ineffective, since the war raging on the French coast isn't redundant at all when it comes to explain the trees' nudity. And, by the same reasoning, it wouldn't help either to follow Haslanger's lead (Haslanger, 2012, p. 317)[20] and require that the social factors' contribution be significant.

[18] It is worth emphasizing that social constructionists typically don't do that – or at least not clearly.
[19] Thanks to John Horden for helpful discussion here.
[20] By this, I do not mean to imply that Haslanger endorses a causal understanding of SOCIALITY AS COOPERATION, but only that she defines causal social constructions in terms of *significant* causal contribution of social factors.

At this stage, we can do two things: keep on looking for an adequate specification of the way in which social factors must contribute to a fact's or event's causal history to be sociality-makers, or drop altogether the idea that some things belong to social reality merely in virtue of causally depending upon social factors. My suggestion is to go for the latter option. Indeed, I doubt that we could come up with a specification that would be intuitively adequate, not too complex and not dangerously ad-hoc. Furthermore, I've already argued (in rejecting (2)) that at least some social things aren't social because of their causal history. Thus, as I've already remarked, accepting the claim that causal dependence upon cooperation sometimes suffices to make for sociality leads to disjunctivism about the social. Everything being equal, it is good to avoid such disjunctivism, and we hence have a pro tanto reason to reject the claim that some things are social merely in virtue of their causal dependence upon cooperation. Finally, I believe that there isn't much pressure to reckon as social any fact (or event) that is socially constructed merely in the causal sense. Upon reflection, it seems perfectly all right to claim that – in the story we used to illustrate the notion of causal social constructions – the fact that blue-eyed Dothrakis are afraid of water isn't itself social, in spite of it social causes; in the same sense, a natural reaction to the question "Do you think that global warming is a social phenomenon?" would appear to be perplexity first and then a question – "You mean that it has social causes?", thus suggesting that using "social" to mean "caused by social factors" is a stretch, albeit one that is certainly intelligible in suitable contexts.

Thus, I conclude that we should not conflate the part of reality which is socially constructed (in the *causal* sense pointed at by Diaz-Leon and Haslanger) and social reality itself. Just like intentional reality has (causal) results which aren't intentional themselves (I intentionally cut an apple into two pieces. Hence, there is an apple cut into two. This fact was intentionally brought about – or intentionally constructed, to pursue the analogy; but it isn't intentional itself.), so does social reality.

1.3.2. Modality: necessitation and supervenience

Next, I shall consider modal[21] specifications of SOCIALITY AS COOPERATION. To begin with, I will wonder whether being social is aptly characterized as follows:

Necessitation: social $=_{def}$ necessitates cooperation.[22]

No doubt, the claim that the social necessitates cooperation is very appealing. That, without cooperation, there could be no laws, no money, no borders, no government, no families, no societies, etc. almost sounds like a truism. Furthermore, it does seem that paradigmatically non-social things (mountains, rivers, isolated human beings, etc.) could exist in the absence of cooperation. Thus, that it necessitates cooperation is arguably not only true, but also distinctive of social reality.[23] But we should pause before taking this observation as decisive. For there is an important difference between **Necessitation** and the extremely plausible claim that the social and only the social necessitates cooperation. The latter merely states that social reality is the unique phenomenon standing in a certain modal relation to cooperation; the former adds that this exhausts the content of the notion of sociality.

To see why we need to pay attention to this difference, consider a thesis that we may label hard necessitarianism. According to hard necessitarianism, everything is necessarily how it actually is.[24] Intuitively, hard necessitarianism *has nothing to do* with social reality. In particular, whether it is true or not is irrelevant to the extension of social reality. Hence, that mountains aren't social is intuitively completely unrelated to hard necessitarianism – and so is the fact that the beginning of the Huronian Ice age wasn't a social event. **Necessitation** clashes with this intuition. For suppose everything necessarily is how it actually is; hence, since there actually is cooperation, everything necessitates

[21] The modality here is metaphysical.
[22] To avoid possible misunderstandings, let me make explicit that (1) **Necessitation** entails: necessarily, if there is something social, then cooperation is instantiated one way or another, (2) **Necessitation** does NOT imply that every particular realization of social reality necessitates a particular cooperative configuration (as opposed to an instance of cooperation whatsoever).
[23] I'm being very swift, but harmlessly so: in Section 1.2.2, I've already given elaborated arguments which support the claims I'm making here.
[24] Hard necessitarianism is the kind of view that some – maybe mistakenly (see (Newlands, 2013)) – attribute to Spinoza.

cooperation. Thus, as per **Necessitation**, everything is social. Hence, **Necessitation** has the extremely counterintuitive consequence that hard necessitarianism isn't irrelevant to the extension of social reality.

As I see it, there are two lessons to be learned here. On the one hand, the truth of the claim that the social and only the social necessitates cooperation (if it is indeed true) reflects the modal structure of reality as much as the nature of social reality. On the other, there is more to the relation between being social and cooperation than a mere pattern of modal covariation. Purely modal accounts such as **Necessitation** fail to reveal an important aspect of this relation: the one which makes it the case that, even supposing the truth of hard necessitarianism, not everything is social.

I want to be clear about the form of the argument I'm making. It is easy to be misled into thinking that I'm arguing as follows: there is a description of reality *worth taking into account*, such that, if it is correct, then **Necessitation** has unpalatable circumstances; therefore, we shouldn't accept **Necessitation**. Clearly, this line of argument is rather weak, for the prospects of vindicating the premise (when the description of reality in question is hard necessitarianism) are very low – hard necessitarianism does not seem worth taking into account, at least not in the present context. But my argument is of a different kind. I'm saying: there is a (admittedly crazy) description of reality whose accuracy is intuitively irrelevant to the question as to whether everything is social or not; but if **Necessitation** were true, the accuracy of this (crazy) description of reality would be very relevant to this issue; thus, we should reject **Necessitation**. And, crucially, this argument doesn't rely on the premise that the (admittedly crazy) description in question is worth taking into account. It only requires that it be intelligible enough.[25]

Someone who wants to stick to a purely modal account of the concept of sociality may then offer the following:

[25] If you think hard necessitarianism isn't intelligible enough, consider the view that cooperation occurs necessarily – as would arguably hold someone believing that there are angels, who exist and cooperate necessarily.

Supervenience: social $=_{def}$ supervenes upon cooperation.

But it is unclear that moving from **Necessitation** to **Supervenience** brings in more benefits than costs.

For, just like **Necessitation**, **Supervenience** clashes with the intuition that hard necessitarianism is irrelevant to the extension of social reality.[26]

Furthermore, there are reasons to doubt that social reality supervenes upon cooperation – let alone that such a modal claim captures the concept of sociality. Consider the following pair of stories:

> *Death in the evening*: Since May 1977, Remy is the Swiss' (a primitive tribe) supreme judge: whenever there is a conflict among the Swiss, Remy is to decide how it should be resolved – and his call is final. On the 27th of November 1983, *at 8 PM*, Remy dies and hereby stops being supreme judge. His death remains unnoticed until the 28th of November, at 9 AM.
>
> *Death in the morning*: Since May 1977, Remy is the Swiss' (a primitive tribe) supreme judge: whenever there is a conflict among the Swiss, Remy is to decide how it should be resolved – and his call is final. On the 28th of November 1983, *at 8:30 AM*, Remy dies and hereby stops being supreme judge. His death remains unnoticed until the 28th of November, at 9 AM.

At least on the face of it, *Death in the evening* and *Death in the morning* could unfold in two cooperatively indistinguishable worlds, *w** and *w***. But then, *w** and *w*** would be cooperative duplicates without being social duplicates: for, on the 27th of November 1983, at 9:00 PM, the Swiss do have a supreme judge in *w**, but not in *w***. Thus, the claim that social reality supervenes upon cooperation may not be on good standing.[27]

---

[26] For those who think that hard necessitarianism isn't intelligible enough: there are other (admittedly strange but arguably intelligible) theses that could play the role of hard necessitarianism in my argument against **Supervenience**. Thus, if there are angels who exist and cooperate necessarily, and if they cooperate between each other in a way responsive to every tiny feature of the non-angelic world (say they cooperate in building a complete representation of the non-angelic world), then everything non-angelic arguably supervenes upon cooperation (there can be no difference in anything non-angelic without there being a difference in how the angels cooperate); and yet, intuitively not everything non-angelic is social – mountains aren't, and neither are rivers and lonely lions

[27] On a similar line, Brian Epstein (Epstein, 2009) has argued for the stronger claim that social properties do not supervene upon individualistic properties (i.e. properties of individual agents); rather, he claims that

To conclude: one way or another, there are plenty of reasons to remain unsatisfied with modal renderings of SOCIALITY AS COOPERATION. In the next section, I'll argue that we'd have better luck if we tried to try to cash out "somehow involves" in terms of essence or nature.

1.3.3. Essence

In order to construe my final account of social reality, I suggest that we appeal to the notion of essence or nature. Following the work of the many philosophers who've been discussing the notion of essence (and related notion) over the past years, I understand this concept as the one in use in the following examples:

- Socrates belongs to the essence of Socrates' singleton (but not the other way around) (see (Fine, 1994))
- Being for cutting belongs to the essence of being a knife
- Being red lies in the nature of being maroon (see (Audi, 2012, p. 695))
- Having a spouse belongs to the essence of being a wife

Under different guises, questions relating to essence, essential dependence and other essential connections have been the main topic of a huge literature over the past years (for some sample examples see (Fine, 1994), (Fine, 2012), (Audi, 2012), (Correia & Schnieder, 2012), (Rosen, 2010), (Jenkins, 2011), (Schaffer, 2012), (Schaffer, 2009), (Thomasson, 2007 Ch.2-3) and (Skiles, 2015)), and there is little consensus on many important questions related to these notions (Is essential dependence (and cognate notions) reflexive, transitive, asymmetric, primitive, fundamental? How

environmental properties must also be included in the supervenience basis. His argument relies on cases like the following. The Swedes (a primitive tribe) have the following rule: their totem – which they have to cherish, protect and worship – is the highest tree in their territory. Contrast now two scenarios. In the first, the Swedes worship tree T, which actually is the highest tree in their territory. In the second, the Swedes still worship T, but – unnoticed to them – it is actually T* which is the highest in their land. According to Epstein, the social property of being the Swedes' totem has a different extension in the two scenarios, although the distribution of individualistic properties may be identical. There is an immediate reply: to deny that the Swedes' totem is the highest tree in their territory and insist that, rather, it is the tree that the Swedes deem the highest, which instantiates the property. But, at least without further elaboration, this reply is unsatisfactory, since it undermines – by analogy – the difference between, for instance, money and fake money (fake banknotes seem to be nothing but notes of which some people think that they meet the conditions for being money). Thus, as it stands, Epstein's argument is another threat the defender of **Supervenience** shall have to dodge.

does essential dependence relate to grounding and metaphysical explanation? What is the relation between essential dependence and conceptual dependence? Is essential dependence fundamental, or is it rather derivative?). It is obviously not the place here to review and discuss this literature. Fortunately, I do not need to do so: for the main part, my discussion will only rely on the intuitive thought – almost common ground among those positively predisposed towards essence talk – that essential connections go beyond modal connections (at least when the modality in question is metaphysical). This intuition is typically motivated by examples like the following:

- Socrates belongs to the essence of Socrates' singleton, but not the other way around (while metaphysical necessity presumably runs both way)
- Everything is such that it metaphysically (and maybe even conceptually) necessitates that nothing is both blue and red all over, but that nothing is both blue and red all over doesn't belong to the essence of everything

Unsurprisingly, some people have expressed skepticism about notions which are supposed to take us beyond metaphysical modality. They typically claim that such notions are either unintelligible or useless (for expressions of skepticism about the cognate notion of grounding see (Wilson, 2014) and (Daly, 2012)). I myself find this skepticism generally ill-motivated and believe that it usually draws on unnecessary assumptions about the functioning and content of the notion of essence (or, for what matters, grounding). But I won't try to convince the skeptics here: I'm likely to fail, since the arguments I have to offer are already on the market. Hence, it is an assumption of mines that there is a notion of essence, deployed in examples such as the above, which express connections that go beyond metaphysical modality.

Our last rendering of SOCIALITY AS COOPERATION henceforth reads:

Essence: social $=_{def}$ involves cooperation essentially

The arguments of Section 1.2.2 constitute the main argument for **Essence**. For, although they are meant to support SOCIALITY AS COOPERATION on a less specific guise, they have arguably exactly the

form they should have to support the claim that cooperation is essentially involved in social reality. This is particularly clear of the discussion of paradigms of social kinds: it focuses on what it is to be money, a leader, a war, an institution, etc. – i.e. it asks about the nature of these kinds – and argues that this nature makes reference to or involves cooperative activities. Hence for instance, I've argued that:

- the nature of money is to be a means of exchange; and exchanging is a cooperative activity
- the nature of war is to be a conflict among *groups* of people, i.e. that is, in war, each side is constituted by many people who *cooperate* with one another
- it belongs to the nature of leadership that the leader and the led cooperate in an action whose distinctive aspect is this: having a leading and a led part.
- it is in the nature of institutions that they are created (partly) by commitments grounded in cooperative activities.

Hence, I take it that the discussion of Section 1.2.2. constitutes an argument not only in favor of the original SOCIALITY AS COOPERATION, but also for its essentialist specification (i.e. **Essence**).

On a different line, the plausibility of the claim that social kinds necessitate cooperation gives us another reason to buy into **Essence**. For – on pain of indulging in many unexplained modal connections – we will have to say why the instantiation of money, marriage, law, border, etc. necessitates cooperation. And **Essence** elegantly points at a general answer: these kinds, together with other social kinds, necessitate cooperation because cooperation belongs to their essence or nature.

Now, a first objection which I need to address is that I cannot both endorse **Essence** and remain neutral with respect to the irreflexivity of the notion of essential involvement. For, as acknowledged several times, cooperation is itself a social phenomenon. Thus, according to **Essence**, cooperation essentially involves cooperation; and hence essential involvement is not irreflexive. I take this to be a correct observation, rather than an objection. For, on the one hand, I find it extremely plausible that essential involvement is reflexive. Claims such as "being a knife essentially involves being a knife", "cooperation

essentially involves cooperation", etc. seem trivial to me (so trivial indeed, that they also look a bit silly). On the other hand, and maybe more importantly for the purposes of this chapter, if it (surprisingly to me) turned out that essential involvement is indeed irreflexive, I could simply reformulate **Essence** in terms of essential involvement*, defined as the reflexive extension of essential involvement.[28]

Another objection I want to dismiss at this point claims that **Essence** makes it too cheap to belong to social reality (i.e. that it overgeneralizes). Here come two potentially worrisome examples:

- Being a desire to cooperate
- Being a situation in which cooperation would be fruitful

Presumably, the objection goes, cooperation belongs to the essence of the two properties denoted above; and yet, none of them is a genuinely social property. Thus, **Essence** is too weak.

To start assuaging this worry, notice that one should not conclude that a property essentially involves cooperation just because it can be referred to by means of a description which involves the notion of cooperation. Otherwise, (almost) any property would essentially involve cooperation (and everything would essentially involve anything); for "being P" and "being P and such that cooperation is cooperation" have the same denotatum, if any. Thus, we should be careful not to conclude that cooperation is essential to the two purported counterexamples above merely in virtue of the constituents of the terms we use to name them.

Bearing this in mind, we can deal with the purported counterexamples by insisting that, under the intended interpretation, essential dependence has modal consequences. Explicitly, we may get off the hook by requiring that, if X is essential to Y, then Y necessitates X. For then, neither situations in which

[28] A similar objection could be raised to the effect that I cannot remain neutral with respect to the transitivity of essential involvement. For instance, the objection would have it that I must say the following: *inflation* is a social phenomenon because money belongs to its nature, and cooperation belongs to the nature of money. I'm not sure that the pressure to accept the transitivity of essential involvement is a strong as it is in the case of irreflexivity (after all, isn't it plausible to say that inflation is, by nature, the devaluation of a means of exchange?). But, in any case, my reply would mirror the one I offer to the irreflexivity quarry.

cooperation would be fruitful nor desires to cooperate essentially involve cooperation, since they do not necessitate it.[29] Given the plausibility (highlighted in Section 1.3.2.) of the claim that the social necessitates cooperation, I do not expect such precisification of our essentialist framework to create any major trouble. Nevertheless, I take it that it would be nicer to deal with the cases at hand (and their likes), without burdening our essentialist framework and our account of sociality with new commitments. I shall try to do this in what follows. The reader who remains unconvinced may always turn back to the solution discussed in the previous lines.

The key idea is to carefully distinguish between essentially involving cooperation (or, more generally, essentially involving BLA), and being individuated – among other things – in a way related to cooperation (or, more generally, being individuated – among other things – in a way related to BLA). While it is quite plausible that the former implies the latter (but you don't have to be convinced – this claim doesn't play any role in my discussion), I take it that the opposite implication fails. Consider the property *being a world with no trace of cooperation*. No doubt, this property is individuated in a way related to cooperation: it is the presence or absence of cooperation that determines whether it is instantiated or not, and any property that follows this cooperation-related-pattern is identical to it.[30] But from this it clearly doesn't follow that this property essentially involves cooperation: quite the opposite, it appears to essentially involve the *absence* of cooperation and – as such – it doesn't involve cooperation *at all* (let alone *essentially*).

If I'm right in the above, and if the intuitive difference I appeal to is robust enough,[31] we can deal with the purported counterexamples as follows. Firstly, it is natural to think that cooperation (the

[29] True enough, there are theories of content which presumably entail that, *somehow surprisingly*, desires to cooperate necessitate cooperation. I'm thinking of some externalist theories of content, which would have it – for instance – that possessing the notion of cooperation requires standing in suitable causal relations to actual instances of the phenomenon. But on such views, I would argue that desires to cooperate turn out to, *somehow surprisingly*, belong to social reality.

[30] At least under a not-to-fine-grained understanding of properties, which has it that properties are individuated by their intensions (their extension in each possible worlds).

[31] If you think the distinction is not robust enough, and that relying on it makes the account liable to an ad-hocness complaint, you can somehow solidify or fix it by explicitly requiring that essentially involving BLA has modal consequences that being individuated in a way related to BLA lacks. This brings us back to my first reply.

phenomenon) isn't involved itself in desires to cooperate – it is rather the idea of cooperation which is; secondly, a situation in which cooperation would be fruitful is typically a situation in which cooperation does not occur, and hence isn't involved either. True enough, both these properties seem to be somehow individuated in a way related to cooperation, but, as previously argued for, this consideration falls short of supporting the claim that they essentially involve cooperation. Our last version of SOCIALITY AS COOPERATION remains thus intact.

Out of the same concerns (avoiding commitment which I deem unnecessary, and keeping the account as flexible as possible), I do not adopt this proposal myself.

Chapter 2: The nature of cooperation

Abstract: cooperation is a ubiquitous phenomenon. Human cooperate on a daily basis and in countless occasions. Furthermore, as argued in Chapter 1, cooperation is the hallmark of the social. In this chapter I present and defend an account of cooperation. According to this account, cooperation occurs when people pursue a goal they have in common, according to a plan they share, and in a state of (roughly) mutual awareness. After presenting the account, I reply to several objections inspired by the literature on collective action.

2.1. Introduction

Cooperation is a ubiquitous phenomenon. Humans cooperate on a daily basis, and in countless occasions: when they talk to each other, when they buy or sell things, when they drive on the right (or on the left, in some countries), when they play games, when they go hiking together, etc. Furthermore, as I've previously argued (see Chapter 1), cooperation is the hallmark of the social. It is thus time that we dig a bit deeper into the nature of this phenomenon.

In this chapter, I offer and defend an account of cooperation. Roughly, the account says that cooperation takes place when some agents pursue a goal they have in common, by following a plan they share, and in a state of mutual awareness. The outline of my discussion is as follows: in Section 2.2, I introduce the phenomenon to be analyzed, that is, cooperation; in Section 2.3, I develop and argue for a particular account thereof; in Section 2.4, I consider and reject some objections to my account.

2.2. Cooperation[32]

Cooperation is a ubiquitous phenomenon. Humans cooperate on a daily basis, and in countless occasions: when they talk to each other, when they buy or sell things, when they drive on the right (or on the left, in some countries), when they play games, when they go hiking together, etc.

Cooperation is many-faced. There are small scales cooperative ventures, like when my brother and I go for a walk together, prepare a mayonnaise together, clean a park together, and paint a house together, etc. (see (Searle, 1990), (Gilbert, 1990), (Tuomela, 2000), (Bratman, 1993), (Kutz, 2000), (Ludwig, 2007), among others). But cooperation can also be massive (see (Shapiro, 2014) and (Kutz, 2000)): thousands of French soldiers cooperated in World War I and millions of people cooperated across the world in the massive protests against war on Iraq in 2003. Cooperation can be voluntary, like (hopefully) when my brother and I play tennis together. But there are also cooperative activities in which people engage willy-nilly: I may decide to cooperate with a dictatorial political regime out of fear for reprisals, rather than political sympathy; and it doesn't take much imagination to picture employees cooperating in projects they abhor in order not to lose their jobs. Finally, while cooperation might result from altruistic motivation, it doesn't have to be so. If Jonas is moving out and asks his friends to come and help, those who cooperate in the moving presumably acted on altruistic motivations. But, if both Paul and Sarah, who are managing partners in a law firm, want to lay off George, the third managing partner, they might decide to cooperate in order to be able to vote him out, and their cooperation is likely to be based on sheer selfish interest.

Cooperation is closely related to the phenomenon of doing things together. Many (if not most, or all) instances of doing things together are instances of cooperation, e.g. going for a walk together, playing tennis together, etc. Furthermore, cooperating is arguably a paradigm of something that agents do together – that is, any instance of cooperation is an instance of doing things together. Nevertheless, it may be argued that these are two different phenomena; that is that agents can do things together

[32] I've already offered these preliminary remarks in Chapter 1. I repeat them here for the sake of readability.

without cooperating in doing it. A simple example thereof maybe environmental contamination: arguably humans contaminate the environment together, and yet they do not cooperate in doing so. Or, for a more complex example, consider the following case: Nora is sad and hungry; I want her to be happy and thus give her a happiness pill; Ronald wants to calm her hunger and feeds her; it turns out that, when taken on a full stomach, happiness pills make you dance until you faint – and so does Nora. Arguably, we (Ronald and I) made Nora dance until she fainted; and this is something we did together (none of us did it on its own). But, clearly, we did not cooperate towards this end. Hence, individual actions sometimes add up to constitute a collective action – something that several agents did together – even though this collective result wasn't intended by anyone.[33] And, intuitively, such un-intentional collective actions do not involve cooperation.

2.3. The nature of cooperation

Cooperation involves agents united by common interest. This sounds (almost?) like a truism. Thus, the notion of common interest or common goal appears to be a good place to start our investigation into the nature of this phenomenon. But, things quickly become more complicated when we try to pinpoint more specifically what it takes for agents to have a goal in common in the relevant sense. Here, my strategy will be to start with an extremely simple-minded notion of common goal, and see how (and whether) it can be complemented or weakened to get a satisfactory account of cooperation.

2.3.1. Common goal

To start with, as I shall use it here, an agent's goals are what they desire (understood very broadly). Thus, if I desire (or want) to stop smoking, that I stop smoking is one of my goals, and vice-versa. This might not correspond to our ordinary notion of goal (I suspect that the ordinary notion is more demanding); but I see no reason to worry about using the notion in a partly stipulative manner.

[33] More on this in (Chant, 2006).

It is then tempting to say that two agents have a common goal whenever they desire the same. Unfortunately, the phrase "desiring the same" is semantically too flexible to offer a satisfactory anchoring point. Thus, if Yoann and I want to eat meat, there is a sense in which we desire the same. But, clearly, desiring the same in this sense doesn't ground common interest; it may be quite the opposite indeed – if there is meat only for one person, these desires of ours push us against each other (see (Smith, Lewis, & Johnston, 1989, p. 115). Alternatively, we may say that agents share a goal whenever they have desires with intensionally equivalent content, i.e. desires that are necessarily satisfied or frustrated together. But that is arguably too strong: suppose I want to go with Yoann to Chelsea's home city, and Yoann wants to go with me to Arsenal's home city (see (Bratman, 2014b, p. 42)).[34] Since London is the home city of these two teams, it would seem that Yoann and I have a common interest. And yet, the content of our desires is arguably not intensionally equivalent: after all, Chelsea could be based in Manchester. It thus seems more reasonable to say that common interest only requires extensional match: two agents hence have a common goal if and only if they have desires which, in the actual world, are either satisfied or frustrated *together*.[35]

Obviously though, common goals alone are way too thin a basis to ground cooperation. As many people, I wanted the Swiss Football Team to win the 2014 World Cup, but – albeit sharing a goal with them – I did not cooperate with the other fans of the Nati[36] in order to make my wish come true. Rather, I stayed at home, watched the games on TV, and complained for half an hour when my team finally lost to Argentina. In a nutshell, I didn't do anything to help bringing about the victory of the Swiss players. Hence, I didn't cooperate with those I was sharing this goal with, for those who participate in a cooperative venture do not merely share a goal, *they also actively pursue it*.

[34] Arsenal and Chelsea are two (English) football teams, based in London.

[35] Admittedly, the right-hand side of this biconditional is ambiguous, and not all the possible readings give a satisfactory characterization of common goals. I believe that it is clear enough which reading is the one that should be adopted. But, just to be on the safe side, I shall make explicit that the right-hand side requires that agent A_1 have a desire D_1, and agent A_2 have a desire D_2, such that, in the actual world, D_1 and D_2 are either both satisfied or both frustrated.

[36] Nickname of the Swiss Footbal Team.

But this is still not enough. Paul wants that Rome be green tomorrow, and so does Stephen. They both actively pursue this shared goal of theirs. But they don't cooperate: they rather act independently from each other, i.e. each pursues his goal on his own (maybe because they don't know that they have a goal in common; maybe because, even though they know that they share a goal, they don't want to pursue it together). On the contrary, in cooperation, the participants not only actively pursue a goal they share – they are also somehow predisposed to pursue it with the other participants. Thus, if I cooperate with an association aiming at the protection of Panda bear, not only do I want Panda bear to be protected, I also want *us* (the association's members, the participants) to protect Panda bears. Or if you and I cooperate with each other in building a fence between my house and yours, not only does each of us want that there be such a fence – both you and I also want that this fence be built by us two, i.e. as a result of actions of mine, and actions of yours. More generally, it thus seems that agents who participate in a cooperative project want *that they* bring about a certain outcome, not only that this result be brought about in any way (albeit they desire this *too*).[37,38]

But isn't this too strong a requirement? Consider. I read a post on Facebook inviting people not to use any electricity next Saturday between 6pm and 7pm, in protest against the rising prices of electricity. I decide to go along – and so do many other people. Hence, we end up cooperating in a protest against the rising prices of electricity. Now, according to the claim above, as I cooperate in this endeavor, I want that *we* stop using electricity. But, here comes the worry, how do I represent *us*?

It wouldn't do to say that I think of us as *those who participate in the protest against the rising prices of electricity*? For, trivially, those who participate in the protest stop using electricity. Hence, the goal that the participants stop using electricity is empty – in pretty much the same sense in which the goal

[37] Importantly, we shouldn't think that my desire that we bring about a certain outcome O requires that I be enthusiastic about *us* doing something; I may be moved exclusively by O and believe that joining forces with you is a good way to achieve it; I may even think that joining forces with you is not a specially good idea, if I happened to also believe, for instance, that deterring you from participating will be too costly. In all these cases, I end up desiring that *we* bring about O, because all things considered I've settled on a way to try reaching the outcome I desire which includes actions of yours.

[38] As I understand the phrase here: I want that the Xs bring about J if I want that J be brought about by actions of the Xs, *intentionally or not*.

that the people who will wear a hat next Monday wear a hat next Monday is. As such, pursuing it doesn't drive me to act in one way rather than another: empty goals do not move people (in the present case, pursuing the goal that the participants stop using electricity doesn't even give me a reason to stop using electricity myself).[39]

Hence, it seems that we must assume that I can refer to those who will participate independently of their future participation. But, here is the thing, in the scenario under consideration, it is very unlikely that I have a desire referring in such a way to the future participants *and only them*: after all, I presumably don't even know who will take part and who will stay aside. What is less implausible, though, is that I have a goal that has built into it reference to a collective which countenances all (or at least most of) the future participants: thus, I might desire that *those who've been reached by the Facebook post* stop using electricity at the right time; or else, I may be pursuing the goal that *the inhabitants of my country stop using electricity* at the right time; in both these cases, one could argue that I desire of the future participants that they indeed end up participating – for each future participant belongs to the collective whose members I want to turn off the electricity next Saturday between 6pm and 7pm.

Let me coin some terminology in order to take stock. On the one hand, I shall refer to the participants (the Xs) in a given cooperative venture as *the cooperating collective*; on the other hand, whenever someone wants that some agents (the Ts) bring about a certain outcome, I shall say that the Ts are the *target collective* of her desire. The conclusion of our previous discussion is that it would be too strong

[39] Buterfill makes a similar point in his *Joint action and Development* (see (Butterfill, 2012, pp. 30–32). His argument is less general though. It relies on cases like the following: "Mia and Sobani are in a crowded space. Each intends to move a table and, thanks to her background knowledge, expects that exactly one other agent intends the same. But neither Mia nor Sobani can identify who else she expects to be involved in moving the table, except trivially as the other table-mover" ((Butterfill, 2012, pp. 31)). Butterfill then rightly remarks that the fact that Mia and Sobani can only identify the other potential table-mover as a potential table-mover prevents them from achieving the kind of on-line coordination that is needed if they are to jointly move the table. This argument isn't fully general though, for the kind of on-line coordination needed in the case at hand isn't necessary in all instances of joint action (or, for what matters, cooperation). Hence, in cases of *prepackaged cooperation*, where "we work out, in advance, what role we will each play in our [effort towards the goal we have in common]" (see (M. E. Bratman, 1992, p. 339)), such on-line coordination isn't required. The Facebook example I'm discussing here fits precisely in this category.

to demand that each cooperator pursues a goal involving the *cooperating collective*; rather, the predisposition to work *with the other participants* should be captured in our account by requiring that each cooperators wants that the common interest be served by the actions of a target collective, which contains enough members of the cooperating collective. Nevertheless, to keep the formulation of the account as simple as possible, I shall just say that some agents share the goal that they bring about a certain outcome whenever they meet these conditions.

Thus, I can now offer a first tentative account of cooperation:

> *View 1.* The Xs cooperate if and only if, for some outcome J, (1) the Xs share the goal that they bring about J, (2) each X pursues the goal that the Xs J.

Before moving on, it is worth highlighting a major difference between these first two clauses of my account of cooperation and the first clause of Bratman's account of shared intention and shared cooperative activity (see (Bratman, 1992) and (Bratman, 1993)). Where I say that those who cooperate share the goal that they J and pursue that goal, Bratman says that they intend that they J. Prima facie, this may appear to be a mere terminological difference. But once we take into account Bratman's conception of intention, we quickly realize that the matter goes much deeper.

Bratman accepts what he calls a settle condition on intention, which roughly requires that, when I intend to Φ, (i) this intention of mine settles whether I Φ and (ii) I believe that. When applied to the kind of individual intentions Bratman uses to build his account of shared intention, this condition has the strange and discomforting consequence that if you and I both have an intention that we J, then the intention of each settles whether we J and "each of us needs to believe that his intention really does settle whether we J" (Bratman, 2014, p. 65).

Quite clearly, the settle condition on intention needs to be understood with a substantive grain of salt (for an extensive discussion of the way how the settle condition can be sensibly understood, see (Bratman, 1999 Section 4)). But the details need not worry us here. For, as a matter of fact, Bratman takes the settle condition on intention (whatever it requires exactly) to motivate the claim that, when

two people cooperate, their respective intentions that they J are persistence dependent in the sense that "each will continue to so intend if, but only if, the other continues to so intend"(Bratman, 2014b, p. 65).[40] Equipped with the persistence dependence condition, Bratman shows how intention that they J can satisfy the settle condition. Suppose that:

> "(a) we each intend that we J (…)
>
> (b) there is persistence interdependence between the intentions of each in (a) (…)
>
> (c) if we do both intend as in (a), then we will J by way of those intentions (…)
>
> if we are in the conditions specified in (a)–(c) then each of our intentions in (a) will settle whether we J in part by way of its support of the intention of the other. In such a situation, my intention that we J (in (a)) supports (by way of (b)) the persistence of your corresponding intention, and these intentions of each of us in favor of our J-ing together lead appropriately to our J-ing (as in (c)). My intention that we J leads to our J-ing, in part by way of its support of your intention that we J, and vice versa. The control my intention has over our J-ing goes in part by way of its support of your intention that we J. And vice versa. So my intention in (a) settles that we J, in part by way of its support of your intention that we J; and vice versa. So given (b) and (c), the intentions in (a) each settle whether we will J.
>
> Suppose then that (a)–(c) are true and this is known by each of us. In knowing (a)–(c), each knows that his intention that we J will appropriately lead to our J-ing in part by way of its support of the other's intention that we J (and thereby of the other's relevant actions)." (Bratman, 2014, p. 65-66)

That is, according to Bratman, the Xs (some agents) can each have an intention that they J without violating the settle condition provided that the Xs' intentions are persistence dependent (and the Xs believe this much). For then, each X's intention settles whether the Xs J in part by way of the support it brings to the other Xs' intentions (and each X believes this much).

But now, there are many cases of cooperative activity where the participants are hardly individually (or distributively) seen as holding intentions which settle whether they achieve the goal they share (and attributed the corresponding belief).[41] This is the case of many massive cooperative activities.[42]Remember, for instance, the case of the Facebook organized protest against the rising prices of electricity. How plausible is the claim that, as I decide to go along, my decision (and the

[40] See also (Bratman, 1999), where Bratman discussed for the first time how his account could accommodate the settle condition on intentions.

[41] This need not be seen as an objection to Bratman, though, since – as he repeatedly states – his aim is merely to provide sufficient (as opposed to sufficient and necessary) conditions (see, for instance, (Bratman, 2014a, p. 333).

[42] In his *Massively shared agency*, Shapiro has forcefully argued that, for this reason, Bratman's account cannot be extended to generally cover instances of massive collective actions (see (Shapiro, 2014, pp. 272–274).

intention which follows it) settles whether we stop using electricity between 6pm and 7pm by supporting the intentions of the other participants. Not very much, I believe. For one thing: none may be aware that I have made this decision – and, in such a situation, it is hard to see how my intention could support anyone's intention in the sense of the persistence dependence condition. For another: my contribution to the overall endeavor is likely to be extremely marginal; accordingly, the support (if any) it brings to the intentions of the other participants shall be extremely marginal too, and thus unfit to ground the claim that, in any sensible sense, my intention settles or control whether we stop using electricity between 6pm and 7pm.

Now, crucially, that some agents each pursue a goal they share (in the sense discussed above) doesn't require that these attitudes of theirs be interdependent in the sense of the persistence dependence condition. Let me bring to the fore one more time the example of the Facebook organized protest, and suppose that, just like me, agents X_1 to X_n decided to go along. According to the account of cooperation I propose here, each of us pursues the goal that we protest against the rising prices of electricity by refraining from using any electric energy between 6pm and 7p (that is: each of us desires that we so protest and takes steps towards the satisfaction of this desire of hers). And we may all do this even though our respective attitudes towards this goal do not mutually support each other (that is: even though my attitude towards our shared goal doesn't contribute to make it the case that X_1's attitude persist, or that X_2's persist, etc. – as it is arguably the case if X_1 to X_n aren't aware that I have this attitude). Hence, whatever settle condition on intending motivates the persistence dependence condition doesn't apply to pursuing a goal. And therefore, an account built on the latter notion promises to cover more ground – and, since it allows that agents make individually marginal contributions to the shared goal they pursue (and see their contribution as marginal), it might actually be able to accommodates cases of massive cooperation.

2.3.2. Common plan

Now, it should be clear that *View 1* doesn't give us sufficient conditions for cooperation yet. Consider: Jasmin and Laila are old friends. They like each other very much. But they haven't been in touch for some years. Jasmin wants to run a bar with Laila and *vice-versa*. They have come to such desires independently from each other. At some point, Jasmin starts thinking more and more about the issue: he finally decides to try contact his friend. Unfortunately, his attempts result in a failure. As for Laila, she also feels more and more the urge to share a bar with Jasmin: but – albeit she tries and tries again – she cannot get in touch with her old buddie. Uncontroversially, Jasmin and Laila meet the first condition set out in *View 1*: both want that they run a bar together. Furthermore, when they try to get back in touch, they both actively pursue this goal of theirs – and hence meet the second condition of *View 1*. And yet: there is no doubt that Jasmin and Laila fail to cooperate with each other.

Obviously, in this case, what prevents Laila and Jasmin to cooperate is that they cannot get in touch with one another. But I believe it would be a mistake to conclude that communication among the participants is a necessary component of cooperation. Sometimes, people cooperate by following pre-established patterns which makes communication unnecessary: when I drive my car, I cooperate with the other drivers with whom I share the goal of avoiding car crashes. Often, this cooperation will involve communication among us – but sometimes it may be enough that we strictly abide by the Traffic Code.

But then: how is it that Jasmin and Laila fail to cooperate? And why is it their being cut off from each other that makes cooperation impossible? I suggest the following: Jasmin and Laila do not cooperate because – albeit they share a goal, and a goal that they both actively pursue – they do not share a plan that would guide and coordinate their efforts. And it is their failure to get back in touch that prevents them to cooperate because it is that which prevents them to design a shared plan.

Here comes a different story to support the same conclusion. Say Roger and Jackson both want that they build a house together. Sadly, they have irreconcilable views about how they should do it. For a

while, they try to convince each other but, at some point, they both get very angry and stop talking to one another. Then Roger has an idea: instead of trying to talk Jackson into changing his mind, he will make Jackson's plan impossible to realize, hoping that Jackson will then come to accept his own plan (i.e. Roger's). Thus, since Jackson thought that it would be a good idea to build a wooden house, Roger burns all the wood there is. *Mutatis mutandis*, Jackson follows the same reasoning. It is clear that – once they stop talking to each other – Roger and Jackson do not cooperate with one another. And yet, they share a goal of the relevant form (that they build a house together) and they both actively pursue it (e.g. Roger pursues the goal he shares with Jackson by burning all the wood there is – although he does this in order to force Jackson to accept his plan towards their common goal). But, as in the case of Jasmin and Laila, they do not share a plan, i.e. a way for them to pursue their shared goal. And this suffices to preclude them from cooperating.

At this point, some comments are in order to try making the talk of shared plan more precise. Firstly, as I shall use the phrase, a plan towards a goal describes a way to pursue or achieve that goal (and that exhausts what a plan *is*). Plans may be very incomplete: my friend Aurélien is been locked up in a garage by some gangsters in Washington. I don't know yet how to get him out. For the time being, my plan is to go to Washington: nothing more.

Secondly, to accept a plan towards a certain goal is to settle on pursuing this goal in the way the plan specifies. Thus, if I accept a plan, I'm decided to perform certain actions specified by the plan, as means to achieve the goal the plan aims at.

Thirdly, while some plans concern only one agent, others concern several people. Thus, my incipient plan to rescue Aurélien is only for myself. But I could as well have a plan for me and Carolina (Aurélien's girlfriend) to rescue him: it could read that I go to Washington while she stays in Europe to coordinate our efforts with those of the European police. As I said above, to accept a plan towards a certain goal is to pursue this goal in the way specified by the plan. It is worth explaining a little bit how I want this notion to apply in the case of plans that concern more than one agent. Consider the above plan for

Carolina and myself: that I accept it certainly means that I follow its instructions, as means to achieve Aurélien's rescue, i.e. that I go to Washington. But there is more to it: I know that I'm only one of the persons concerned by the plan and I know that Aurélien's rescue also depends on Carolina's performance. Thus, not only do I settle on doing my share, I hope that Carolina will do hers. Generalizing, when I accept a plan for several people, not only do I intentionally perform the actions it specifies *for me* in order to achieve the goal it aims at; I also hope, desire or want that the other people concerned by the plan successfully do their part.[43]

Fourthly, there are two relevantly different kinds of plans for several agents. Some of them are individual-specific, so to speak; some of them aren't. An example of the latter type would be a plan for my brother and I to bake a cake together saying that, first, we should buy the ingredients and, second, we should use them following the instructions of the recipe. An example of the former type would say that, in order that my brother and I bake a cake together, I shall buy the ingredients, while he warms up the oven, and then we shall meet at our place where he will prepare the doe while I cut the chocolate into small pieces. If plans are to ensure that those who cooperate are somehow coordinated with one another, it is individual-specific plans that are needed: we may agree that, in order to paint our house anew, we need to buy the painting first; still, if we don't agree on who is to do what in order for us to buy the painting, cooperation won't be able to get off the ground. Henceforth, then, "plan" will by default refer to individual-specific plans for several agents.[44]

Finally, I shall say that several people share a plan for them to J if and only if there is a plan for them to J that each of them accepts. Thus, in the above story, Carolina and I share a plan (for us) to rescue

[43] Importantly, what accepting a plan for several agents does NOT requires is that I settle whether the other participants will do their part or not. Neither does it require that I consider myself as somehow controlling whether the other participants will do their share or not. In this sense, accepting a plan for the Xs to J is still significantly weaker than intending that the Xs J according to a given plan in Bratman's sense (see (Bratman, 2014b, pp. 63–66)).

[44] It is worth underwriting that, as I intend the expression, individual-specific plans do not need to make direct reference to individuals (as do proper names such as "Aurélien", "Carolina", etc.). Rather, they need only contain enough information so that those who accept the plan may get to know their respective shares. To this effect, I take it that it suffices that they contain: (1) a description of the shares, (2) for each share, an (individuating) description of the people in charge thereof.

Aurélien. The plan says that I'm to go to Washington and she is to stay in Europe to coordinate with the local police, and we both intend to behave in conformance with it, while hoping that the other will successfully perform her task.[45,46]

To sum up, our account of cooperation now reads as follow:

> *View 2*. The Xs cooperate if and only if, for some outcome J, (1) the Xs share the goal that they bring about J, (2) the Xs share a plan for them to J,[47] (3) each X pursues the Xs' goal, by following the Xs' plan.

[45] As pointed out by Shapiro (see (Shapiro, 2014, p. 279)), we should be careful not to require that the agents who share a plan must have explicit knowledge of all the plan's parts (e.g. the volunteers of an association aiming at preserving panda bear from extinction share a plan, but many of them presumably ignore parts of the plan which do not concern them – those who do field work may not know what the plan requires from those responsible of communication). Fortunately, I don't think that anything I've said so far conflicts this observation: that some people accept the same plan only requires that they be somehow capable of referring to the plan in question (as maybe, the association's plan, the Facebook activists' plan, or the plan inherited from the tradition), not that they have explicit knowledge of every part of it.

[46] In his *Massively Shared Agency* (Shapiro, 2014), Shapiro develops a closely related notion of shared plan. According to his proposal, "a plan is shared by a group to J when (1) the plan was designed, at least in part, for the members of the group so that they may engage in the joint activity J and (2) each member accepts the plan" (Shapiro, 2014, p. 278). His understanding of acceptance of a plan is very similar to mine (see (Shapiro, 2014, p. 279). Hence, the main differences hinge on clause (1). *Firstly*, it requires that the goal of the plan be a joint activity, where a joint activity is an "integrated whole that has actions as its parts", as opposed to a mere collection of actions (see (Shapiro, 2014 footnote 35)). I see no reason to impose such a requirement (whatever its content exactly is). Hence, I believe Amya and I could cooperate so that she gets to go to the dentist and I can stay home and cook dinner. In this case, we would share a plan to this end, which looks pretty much like a mere collection of actions to me. Furthermore, I suspect that this requirement makes Shapiro's account circular and, even though, as argued at length in Chapter 4, I do not believe that circularity is a no-go, I take it that it is good to avoid it if possible. *Secondly*, it requires that shared plan be designed, at least in part, for the members of the group. I think this is too demanding. For I take it some plans may emerge unintentionally - and hence without being aptly described as designed for any group of cooperators. Consider, for instance, one of Lewis's favorite examples of convention (Lewis, 1969, p. 52): in a North-American town, all local phone calls are systematically cut off without warning after three minutes; a convention soon emerges according to which, when this happens, the original caller is to call back, while the called party waits. By assumption, the convention wasn't designed for anyone – it is an unintended result of the way how people tried to deal with the situation. Now, suppose Jean and Shura are cut off in the midst of an intense conversation. Jean, the original caller, calls back while Shura waits. As far as I see it, Jean and Shura cooperate in restoring their call – and the plan they share to this effect is the convention in question, a plan which wasn't designed, not even in part, so that they could restore their call.

[47] Inspired by Bratman's classical discussion, some might object that requiring that the cooperators accept "the same plan" is too demanding – rather, all that is needed is that the plans they accept be compatible (or mesh, in Bratman's terminology). *Prima facie*, the point seems well taken: if we are painting a house together and our plan says that you are to buy the brushes, you may plan to buy them in the *Happy Brush*, and I may have no opinion on the issue. Thus, the argument goes, we don't have *the same plan*: but it doesn't matter, for our plans mesh. Still, I believe that going for a meshing plans requirement makes formulation more cumbersome and unnecessarily so: for if our plans are compatible, *we have the same plan*, provided that we look at it at a suitable level of generality.

2.3.3. A sense of sharedness

But *View 2* is still insufficient. This is so because, as it stands, it doesn't require that the participants have a sense of the fact that they *are on the same side* and such awareness is an essential element in cooperation. Here comes an illustration.

> In a room, they are two people, Rodolfo and Teodolfo, and two buttons (a red and a green one). The two guys aren't allowed to talk to each other and, because of some magic trick, they cannot see each other. It is common knowledge between them that: (1) if the red and the green buttons are pushed simultaneously, a clown shows up and sings a song, (2) if you push the red button at five, you get a candy, and (3) if you push the green button at five, you get a puppet. Both want to see the clown singing, but none of them knows that they share this goal. Rodolfo believes that Teodolfo wants a candy; he thus decides to push the green button at five, so that he can see the clown. As for Teodolfo, he believes that Rodolfo is fond of puppets, he thus decides to push the red button at five, so that he gets to see the clown. They proceed as planned, and the clown shows up and sings.

Now, as far as I can tell, Rodolfo and Teodolfo do not cooperate as they make the clown showing up. They do not cooperate precisely for the reason alluded to before: they aren't aware of sharing a goal; for all Rodolfo knows, the plan he follows (which is a plan *for Rodolfo and Teodolfo* to bring it about that the clown shows up) is aiming at his own satisfaction only; and the same is true of Teodolfo. On the other hand, it would seem that all the conditions set out in *View 2* are satisfied: both Rodolfo and Teodolfo want that they bring it about that the clown shows up; both of them furthermore accept the same plan towards this goal; finally, they both pursue their goal by following the plan they share. *View 2* hence needs to be amended: sharing both a goal and a plan isn't sufficient – some awareness thereof must be required too.[48]

[48] (Blomberg, 2016) contains a detailed and illuminating discussion of the sense of sharedness at stake in this section.

Should we then require that the agents who cooperate with one another believe that they share both a goal and a plan? Albeit this may be the most straightforward implementation of the sense of sharedness requirement this raises several potentially worrisome issues that I need to discuss.

Firstly, it may be argued (and has indeed been claimed in (Alonso, 2009)) that *belief* isn't the right kind of attitude to plug in here. Consider, for instance, the following story (inspired from one of the cases brought up by Alonso):

> Isabela wants to hack her way into the Pentagon's informatics system. She knows that she can't do it on her own. As she tries to break through the many firewalls, she needs someone else to attack the system from a different angle. She meets with her friend Leila, explains her plan and begs for her help. Leila isn't much convinced: although she does think that the plan might work, she's also very much worried about the possible reprisals. She tells Isabela she will think about it. Isabela tell her friend that, in case she decides to join in, she should start the next day at 7PM. Furthermore, for the sake of security, they shouldn't be in touch until then. When the time comes, Leila finally settles on helping her friend hacking her way into the Pentagon's informatics system.

As far as I can tell, this is a story in which two friends end up cooperating to hack the Pentagon's informatics system. And yet, it is unclear that Isabela genuinely *believes* that Leila and she share a goal or a plan. Of course, she might – if, for instance, Leila is generally incapable of refusing to help her (and if she knows that). But the details of the story can be filled in differently: it may be common ground among the two friends that Leila is very cautious and not easily swung; in such circumstances, one may argue that Isabela decides to go on with her plan without properly *believing* that Leila will join in; rather she hopes that she will and, more importantly, she rely on her doing so.[49]

[49] In such circumstances, if Isabela is to be rational, hacking the Pentagon must be very important to her; and Leila must be her only option.

The nature of reliance has been explored to some extent in (Alonso, 2009) and (Baier, 1986). For our present purposes, a few preliminary remarks should suffice. On the one hand, reliance belongs to the family of cognitive attitudes (as opposed to conative mental states such as desire); that is, when I rely on something, I somehow describe the world as being in such-and-such a way (as opposed to prescribing that it should be in such-and-such a way) (see (Alonso, 2009, p. 454). On the other hand, reliance is grounded and justified not only by the evidence the agent has access to, but also by her interests and other pragmatic factors. Thus, in the above story, even though Isabela hasn't enough evidence to properly believe that Leila will join in, she may still rely on Leila's help (if, say, Leila is the only one who could help, and if she – Isabella – *really* wants to break into the Pentagon's informatics system). This is not to say that evidence doesn't constraint reliance at all. As pointed out by Alonso (Alonso, 2009), I may not rely on X being the case if I have conclusive evidence that X is not the case. Hence, if I know (or believe that I know) that the wooden bridge in front of me will break if I try walking across it, I won't rely on it to cross the river (or a least not rationally so); and if Isabela had had conclusive evidence that Leila wouldn't join in and help, she couldn't have relied on her either.[50]

This being said, I think that whenever an agent relies on X being the case, it is also appropriate to describe her as somehow believing that X is the case. In other words, I take the ordinary notion of belief to be broad and flexible enough to cover cases of cognitive attitudes motivated partly by epistemic and partly by pragmatic reasons. Furthermore, I wouldn't be surprised if some other

[50] Appealing to 'reliance' instead of full-fledged belief also allows dealing with an example imagined by Dirk Ludwig:

> "Suppose that country X launches a pre-emptive nuclear strike against country Y. After the initial strike, some missile silos in country Y are still operative. However, country Y has established an elaborate procedure for firing its missiles as a safeguard, which requires two on-site operators, who are physically isolated from one another, and one remote operator, all to punch in a secret code and turn a firing key at their locations in order to launch a missile. Consider the team charged with this for surviving silo 451. After the strike, which interrupts communications between them, none of them knows whether the others have survived, and have some reason, perhaps even preponderant reason, to think that they have not. Nonetheless, they intend to launch the missile. Each of them intends that they do it, and so each of them intends to do his part in launching the missile. Each punches in his code, and then turns his key, hoping that there are still others who are doing their parts, however unlikely it may seem; and so they launch the missile in silo 451 together, and they do so intentionally." (Ludwig, 2007, p. 388)

scenarios suggested that, after all, it is sometimes another attitude of the belief family (e.g. assuming or accepting (see (Engel, 1998))) which grounds the sense of sharedness characteristic of cooperation. And, here too, I would insist that the ordinary notion of belief is flexible enough to be appropriately applied. Hence, I prefer to stick to the original proposal and require that those who cooperate (somehow) believe that they share a goal and a plan. This excursus was no pointless distraction, though, for it makes explicit how broadly *belief* is understood here and, in particular, that the notion is used in a way that allows for pragmatic considerations to bear heavily upon an agent's belief system.[51]

A second question is related to the *target* of the sense of sharedness. In other words: who do the cooperators believe that they share a goal and a plan with? Remember the example we brought up in Section 2.3.1: I see a Facebook post enjoining people to protest again the rising prices of electricity; I decide to go along and cooperate in the protest. Remember also that this example led us to distinguish between the *cooperating collective* (the people who actually end up cooperating) and the *target collective* (for each cooperator, some Ts are such that she wants the Ts to join in; the Ts are then the target collective of her desire). Given this distinction, it is nothing but natural to wonder whether the target of the 'sense of sharedness' is the cooperating collective, or the target collective. Now, for the same reasons brought up in Section 2.3.1, it would be a mistake to require that each cooperator believe that the cooperating collective share a goal and a plan: for this condition is either empty, or it presupposes that each cooperator knows who will join in and who won't. Thus, it is indeed the target collective that the cooperators believe they share a goal and a plan with. But if, in the above example, we suppose that my target collective is the people reached by the Facebook post, this claim generates discomfort. For it would certainly be utterly irrational for me to believe (even in the relaxed sense in which we are using the term here) that all and every people that were reached by the Facebook post will participate in the protest. Hence, it seems that we should only require that each cooperator believes that enough, or some, or many, or most members of the target collective share their goal, and

[51] It goes without saying that I also use "belief" simply to cover cases of merely dispositional beliefs, whatever those are exactly.

their plan towards it, depending on the particulars of the situation. In the Facebook example, I may be willing to protest merely to express my dissatisfaction, and without expecting that the protest achieves any particular result: then I can rationally decides to join in even though I only believe that a few people will take part; on the other hand, if I want to protest in order to have the electric company changing its prices policy so that I pay less for the electricity I consume, it seems that my participation requires that I believe that enough people will join in, where enough is what is needed to put the electricity company under pressure. I shan't make these details explicit in the statement of the account, in order not to make it become too cumbersome. But the reader shall keep in mind that the clause "each cooperator believes that the cooperators share her goal and her plan towards it" is to be understood flexibly, and in particular as satisfied when each cooperator believes that an adequate number of her target collective share her goal and her plan towards it.

A third issue (also raised and addressed in (Alonso, 2009)) is whether we should only require that the cooperators rely on each other's having the relevant attitudes, or whether we ought to demand also that they believe that they will be somehow successful in carrying them through. Reflection on the case of individual action proves decisive here. If I want to go to the supermarket and I act on this desire of mine, I somehow believe it to be possible that I succeed. In the same sense, if I want that my brother and I go to the supermarket together and if I act on this desire of mine, I somehow believe it to be possible that we succeed. Thus, even though cooperators do not need to have conclusive evidence that they will successfully carry through their shared goals and plans, they should at least not have conclusive evidence that they won't. That is, using the notion of reliance previously sketched, it seems appropriate to require that the cooperators rely on each other's success (or, broadly speaking, believe in their respective success).

Last but not least, one may claim that the sense of sharedness present in cooperation requires not a simple first-order belief, but the whole hierarchy of common knowledge. Indeed, it is customary for

accounts of collective phenomena (e.g. shared agency (Bratman, 2014), collective action (Searle, 1990), or acting together ((Kutz, 2000) & (Gilbert, 1990))) to include a clause requiring common knowledge.[52]

Now, to start with, I believe that we should probably refrain from involving *knowledge* in our account of cooperation (and other forms of collective behavior), on pain of seeing epistemological issues cropping up inopportunely. Say epistemological skeptics are right: humans don't know anything. Does that mean that we don't cooperate? I don't think so. Say epistemic contextualism is right: the truth of knowledge ascriptions depends on contextually fixed epistemic standards. Does that entail that the truth of "Alfred and Jo are cooperating" depends on how demanding our epistemic standard is? That seems preposterous. Indeed, Lewis himself has recognized that what he first labelled "common knowledge" may not involve knowledge at all (see (Lewis, 1978)).[53] This being said, we could still include a clause requiring not only first order beliefs, but rather what Lewis has called 'overt belief' (Lewis, 1978, p. 44), i.e. (i) that each X believe that (1) and (2) holds, (ii) that each X believe that (i) holds, etc. And, I believe that we should. Consider:

> In a room, they are two people, Rodolfo and Teodolfo, and two buttons (a red and a green one). The two guys aren't allowed to talk to each other and, because of some magic trick, they cannot see each other. It is common knowledge between them that: (1) if the red and the green buttons are pushed simultaneously, a clown shows up and sings a song, (2) if you push the red button at five, you get a candy, and (3) if you push the green button at five, you get a puppet. When they entered the room, Rodolfo was told: "both of you win if the clown shows up and sings; but Teodolfo doesn't know that. Rather, he thinks that you win if you get a candy"; as for Teodolfo, he was told: "both of you win if the clown shows up and sing; but Rodolfo doesn't know that. Rather, he thinks that you win if you get a puppet". Rodolfo reasons as follows: "since Teodolfo thinks I want to get a candy, he will predict that I'll push the red button at five; thus, in order to make the clown show up and sing, he will press the

[52] On the notion of common knowledge, see (Lewis, 1969) or (Gilbert, 1992).
[53] Thanks to John Horden here. Both for the argument and the reference.

green button at five. Hence, I shall press the red button at five." *Mutatis mutandis,* Teodolfo goes through the same reasoning. They proceed as planned, and the clown shows up and sings.

I take it that, in this last story, it is at best unclear that Teodolfo and Rodolfo cooperate in bringing it about that the clown shows up and sings. Rodolfo and Teodolfo don't see each other as intending to cooperate – but rather as intending to exploit or take advantage of the other's behavior. And this way of perceiving each other seems at odds with the claim that they are cooperating. Granted, the intuition isn't as clear here as in the cases discussed earlier; but, I believe, it still suggests that the presence of higher-order beliefs that rule out such scenarios does belong to the core of the phenomenon we are after.[54]

Furthermore, I think that there is a more general argument to the conclusion that cooperation requires overt belief. Shortly put, we should include the overt belief condition, because it follows from the conceptual truth that cooperation is self-referential. Unpacking the argument a little: it belongs to the content of the notion cooperation that those who cooperate believe that they are cooperating; and the overt belief condition is implied by this aspect of cooperation.

The second premise is unassailable: suppose the Xs cooperate; then, given the self-referentiality of cooperation, each X believes that the Xs cooperate; hence, if, as a matter of conceptual truth, cooperation requires that a certain condition obtains, each X believes that this condition obtains; but then, since it is a conceptual truth that cooperation is self-referential, each X also believes that each X believes that the Xs cooperate; and so on.

Admittedly, though, the first premise is more controversial. Still we should accept it: for none could ever be surprised to be part of a cooperative venture in the same way one can be surprised by having

[54] Admittedly, this scenario only motivates the inclusion of second order beliefs. I won't try to motivate the inclusion of higher-order beliefs by appealing to possible scenarios, though, for matters immediately become too complex to serve as efficient intuition pumps. Hence, I rely on the thought – plausible enough, I take it – that if the way how cooperators must perceive each other motivates the inclusion of second order beliefs, it shall also motivate the inclusion of yet higher-order ones.

performed an action she didn't believe she was performing. My brother and I are playing together. Our game is to throw stones at a wall, aiming at the center of a circle we drew there. At some point, I throw a particularly heavy stone and the wall breaks down. I broke down the wall, and this is quite a surprise. I take it that there is no way that anyone could ever be surprised like this to find herself cooperating. Granted, cooperation can be surprising in many ways. I can be surprised when I find out the identity of my cooperators. And, to some extent, I can be surprised by the numbers of them (remember the Facebook based protest example). Finally, I can be surprised by the success of a cooperative endeavor. But these possibilities of surprised are covered by (i) the fact that I need only one description of the target collective (a description which, in one sense, may fail to reveal the identity of the members of such collective), (ii) the fact that cooperators must only believe that enough, or some, or many members of the target collective will join in, (iii) the fact that the account I put forward doesn't require that the participants be very confident about the success of their enterprise. Thus, they do not undermine the claim that those who cooperate believe that cooperation is taking place.

I can think of one main objection to the above argument. Some shall complain that my argument equivocates. The first premise, they will say, is ambiguous between (at least) two readings. On the first, each cooperator must believe that *she* is cooperating. On the second, each cooperator must believe that *the cooperators* are cooperating. The objection then goes on by claiming that while it is only the second reading which yields a sound argument, it is only the first reading that is supported by the reasoning I've offered. To this I reply that, albeit the first premise can indeed be understood in two different ways, this distinction is irrelevant here: for, as a matter of conceptual truth, I can't cooperate *alone* – thus, if I believe that I'm cooperating, I believe that I'm cooperating with other people who are cooperating.

In his *Common Knowledge and Reductionism about Shared Agency* (Blomberg, 2015), Blomberg suggests that there are relatively clear and mundane counterexamples to the claim that intentional joint action must involve common knowledge (or for what matters, overt belief).

> "consider any type of joint activity that can also be performed by a singular individual, such as going for a walk or making a hollandaise sauce. In such a joint activity, one party may falsely believe that the other mistakenly thinks that she herself intends to carry out the activity whether or not the other joins her (so that the satisfaction of the intention is compatible with the other's involvement but doesn't require it). I may falsely believe that you are under the mistaken impression that I simply intend to go for a walk, rather than that I intend that we go for a walk (that is, what we are supposing that each actually intends). Such false higher-order beliefs are arguably a common upshot of insecurities and mild forms of paranoia that are often present in human relations. And such false higher-order beliefs and doubts can persist throughout joint activities that at least appear to be jointly intentional. But if our walking is jointly intentional only if the CK-condition is satisfied, then such appearances must be illusory. However, in all such cases, agents do not merely each intend to perform individual actions that accidentally have a joint effect. Rather, each intends that they enact the whole action. Due to the interdependence of their intentions, they each settle that they enact the joint performance. Each is thus responsible for bringing it about. Why isn't such a joint performance then an intentional joint action?" (Blomberg, 2015, p. 5)[55]

For all I have said so far, the notion of intentional joint action Blomberg has in mind (e.g. the notion of intentionally doing something together) may well be the notion of cooperation: the examples of non-cooperative joint activities (e.g contaminate the environment) put forward in Section 2.2 are all cases of non-intentional joint actions.[56] The matter is more complex though. On the one hand, some may claim (maybe partly stipulatively) that intentional joint action requires that the participants hold Bratmanian intentions that the joint action be performed, in which case I would argue that cooperation is, in this respect, less demanding (see Section 2.3.1, pp. 45-47). On the other hand, some may claim that intentional joint action is not essentially intentional (i.e. that people may engage in intentional joint actions without intending them *qua* intentional joint actions and hence without believing that they engage in an intentional joint action (see (Blomberg, 2015, p. 9)) – in which case I would argue that cooperation is, in this respect, more demanding (see pp. 60-61 above). This doesn't matter here, though: clearly, the cases Blomberg brings to the fore (we go for a walk together and, as we proceed, my lack of self-insurance and mild paranoias have me thinking that you don't think that I really want to go for a walk with you, but would rather do it on my own), can be seen as instances of cooperative activities. Hence, if his argument is on the right track, we should reject the overt belief condition. But it isn't.

[55] (Blomberg, 2015) also puts forward a case very similar to the second version of the Teodolfo and Rodolfo story. He then suggests that the scenario plausibly depicts a case of joint intentional action. For all I claim, he might be right on this – for as acknowledged below, the notion of joint intentional action may not be identical to the notion of cooperation; but I see no pressure to accept the claim that his scenario depicts a case of cooperation.
[56] On non-intentional joint action, see (Chant, 2007).

To start with, observe that the "insecurities and mild form of paranoia" Blomberg appeals to would – if they had the effect Blomberg presumes they have – undermine not only the claim that overt belief is required for cooperation (or intentional joint action), but also the claim that those who cooperate (or those who engage in intentional joint actions) must believe that they share a goal. For just like we can go for a walk together even though my insecurities and mild paranoias have me thinking that you don't think that I really want to go for a walk with you, we presumably can go for a walk together even though my insecurities and mild paranoias have me thinking that you don't really want to go for a walk with me.[57] This, I take it, should make us pause and put to question Blomberg's understanding of the situations he describes.

Consider the following scenario: there is a huge rock in front of Joe's door. Joe is moving it. Mohammed sees Joe and come to help. They finally cooperate in moving the rock. In this scenario, Mohammed presumably believes (not even mistakenly) that Joe would have "carried out the activity (i.e. move the rock) whether or not he joins him" and, presumably, Joe is aware that Mohammed believes this much. Does this undermine in any sense the claim that, Mohammed also believes that Joe wants (or intends) that they move the rock together (and that Joe is aware of that?). I don't think so. Pretty obviously, one can intend to do something with or without the help of others and nevertheless appreciate a helping hand. And thus, pretty obviously, one can believe that; and believe that someone believes that; and so on.

Consider yet another scenario: Maria and I are going for a walk together. I'm a little paranoid and (mistakenly, this time) think that Maria would rather have gone for a walk on her own. She is aware of that. Does this mean that, as we walk together, (1) I do not believe that she has decided to go for a walk with me (and therefore intends that we go for a walk together); (2) she doesn't believe that I'm aware that she has decided to go for a walk with me. I don't think so. Pretty obviously, one can decide

[57] This is, I take it, quite problematic for Blomberg, since he definitely wants to stick to the claim that those who participate in an intentional joint action believe that they share a goal (see (Blomberg, 2016), and (Blomberg, 2015 footnote 7).

to do something with someone else without being enthusiastic about it. And thus, pretty obviously, one can believe that; and believes that someone believes that; and so on.

In both these scenarios, the kind of beliefs, insecurities and mild paranoias that Blomberg take to undermine common knowledge (or, for what matters, overt belief) are present, and yet, plausibly enough, so is common knowledge. The following diagnosis therefore suggests itself: the intuition that cooperative activities can unfold in the presence of such insecurities and mild paranoias doesn't undermine the overt belief condition. For, these unpleasant states of minds can be plausibly seen as targeting what motivates the participant to adopt the shared goal, rather than the fact that they have such a shared goal (hence, in the second example, I do not doubt that María pursues the goal that we go for a walk together, but rather that she does this for the right reasons: I may think that she does this out of pity, rather than out of the desire to spend time with me; and María knows that I believe that she has decided to go for a walk with me; but she also believes that I think that she hasn't made this decision for reasons I would be delighted with).[58] Hence, I take it that Blomberg's cases give us no reason to believe that cooperation could unfold in the absence of overt belief.

I thus tentatively conclude that we should require overt belief in our account of cooperation. This said, the argument I've put forward in support of this claim raises an interesting question. I have argued that, as a matter of conceptual truth, cooperation requires that the cooperators believe that cooperation is taking place. We thus may want to know whether, beside the overt belief condition, we should add an extra requirement, making explicit this self-referential aspect of cooperation. And some may worry that such an extra clause would doom our account, on the ground of unacceptable circularity (in Chapter 4, I argue that such a circularity would anyway not be sufficient reason to dismiss the account; nevertheless, as I also discuss in Chapter 4, it is reasonable to try avoiding circularity as far as possible).

[58] Of course, insecurities and mild paranoias can also sometimes undermine beliefs regarding the shared goal (first and higher-order). But for all Blomberg says, we have no reason to accept that, in these cases too, intuitions to the effect that cooperation takes place stay strong.

I do not think that such an addendum is necessary. For it is presumably in the nature of overt belief to induce this kind of self-referentiality.[59] That is, whenever it is overtly believed in P that A, then P's members also believe that it is overtly believed in P that A. This claim has been exhaustively discussed and defended by Lewis (Lewis, 1969, pp. 60–68) and Gilbert (Gilbert, 1992, pp. 193–195). It isn't the place here to discuss their arguments in detail. But here comes, at least, a rough motivation. Whenever something P is overtly believed in a given population, it is *unreasonable* to expect that, for each natural number n, the members of this population have a particular reason to hold a higher-order belief of degree n with respect to P (that is, a reason different for the one supporting higher-order beliefs of degree n+1 or n-1, for instance). Rather, they must have a single generic reason supporting all the scales of the hierarchy. That is, something R must be such that, for whatever natural number n, it supports the claim that the corresponding higher-order beliefs are in place. But then, R shall also support the claim that, for whatever natural number n, the corresponding higher-order beliefs are in place. And this is just to say that R supports the claim that the whole hierarchy of higher-order belief holds, i.e. the claim that P is overtly believed.

I thus conclude this discussion of the sense of sharedness involved in cooperation by proposing a third account (which should be understood with the grains of salt previously discussed):

> *View 3*. The Xs cooperate if and only if, for some J, (1) the Xs share the goal that they bring about J, (2) the Xs share a plan for them to J, (3) each X pursues the Xs' goal, by following the Xs' plan, and (4) it is overtly believed among the Xs that (1)-(3) hold.[60]

[59] And hence, overt belief (and common knowledge) allows us to get self-referentiality without circularity.

[60] In *What is cooperation?* (Tuomela, 1993), Tuomela argues that (1) there are cooperative and non-cooperative joint action types (an example of the former being playing a non-competitive game, whereas an example of the latter may be playing chess), (2) joint actions can be performed with a more or less cooperative attitude. As for (1), cooperative joint action types are those in which participants have, at least in principle, reasons to help each other performing their respective parts as well as possible (whereas this isn't the case with non-cooperative joint action types). As for (2), joint actions are performed with a cooperative attitude insofar as the participants are assumed to be disposed to help each other performing their respective parts (see (Tuomela, 1993, p. 96)). One may think that (1) conflicts with the account of cooperation I've put forward here: for, according to *View 3*, any instance of playing chess must be an instance of cooperation. But I take it that this need not be the case. Firstly, Tuomela himself grants that instances of non-cooperative action type must have a "cooperative" bottom (Tuomela, 1993, p. 92) (he then goes on to say that "cooperative" has here a different sense, but doesn't offer

2.3.4. Some illustrations

In order to make sure that the account is properly understood, I shall now put it to work with some illustrations.

Scenario 1:

> Bob's car has just stalled on a hill. It won't start again without a little help. Malorie, Jonathan, and Ainara, upon seeing poor Bob in his predicament, decide to go and help him. Malorie and Ainara push the car from behind; Bob pushes it from its side. Jonathan is on the driving seat, making sure the engine does take advantage of the push to get started.

Here comes a very natural way to understand this scenario. Bob, Malorie, Ainara and Jonathan all want to get Bob's car started. Furthermore, they share a plan to achieve this goal: Malorie, Ainara and Bob are to push (from different places), while Jonathan is to steer. Thus, in conformance with the intuitive verdict, my account says that Bob, Malorie, Ainara and Jonathan do cooperate in Scenario 1.

Scenario 2:

---

any argument to back up this claim). Secondly, my account has the resources to account for the intuitively compelling distinction (which should clearly be seen as gradual, rather than categorical) Tuomela brings to the fore. The idea is simply that some joint action types motivate a more or less thorough cooperation (in the sense of *View 3*) among their participants. If we play chess together, we must cooperate to set up the framework of our game – but, at least in usual circumstances, we have no reason to cooperate further (hence, I have no reason to cooperate with you in order that you find the best moves you can play). On the other hand, if we are doing an escape room together then, not only do we have reason to cooperate in trying to find the way out of the room we willingly trapped ourselves in, we also normally have reason to cooperate in helping each other do our respective parts (e.g. solving an enigma, slipping through a narrow tunnel, etc.). And the same kind of reasoning explains the distinction (once again gradual rather than categorical) that is pointed at in (2). Does this mean that my account and Tuomela's are identical? This is a hard question. Tuomela carefully distinguishes many types of cooperative activities. For instance, in (Tuomela, 1993), he distinguishes between cooperative joint actions (where a joint action is based on a "shared we-intention about which there is mutual belief" (Tuomela, 1993, p. 88)), and cooperative coactions (where coaction is a "collective action in which agents without having a joint intention, have the same goal, perhaps mutually believing so and possibly interacting in various ways" (Tuomela, 1993, p.88)); and, in (Tuomela, 2016), he "mostly concentrate[s] on cooperation as activity based on and involving joint action and especially on joint action `as a group', a strong kind of cooperation" (Tuomela, 2016, p.65). In this situation, I take it that finding out which of Tuomela's account of cooperation (if any) is the one that should be compared to mine is, in itself, a difficult and lengthy philosophical endeavor which shouldn't be undertaken here.

> Anita and Roberta are warladies. They lead two hordes of fearsome warriors. They both want to take Castle Black. They plan an attack together. Anita's horde is to come from the North with battering ram and catapults. Roberta's is to come from the South, and take advantage of the diversion to climb on Castle Black's walls. They follow their plan and win a quick victory.

In this scenario, it is natural to think that Anita, Roberta and their fearsome warriors all want to take Castle Black (more explicitly: that they take Castle Black). Admittedly, they are likely to have myriads of different motivations for this goal – running from the fear of reprisals to the hatred for the lord of Castle Black, passing by loyalty to the warladies –; but this is no obstacle to a common goal attribution. Furthermore, it is plain that the two warladies – and consequently the hordes which obey to them – share a plan to achieve this goal. Thus, my account predicts that the assault on Castle Black shall qualify as a cooperative venture, as it intuitively does.

> *Scenario 3:*
>
> Rob and Octavia are thieves. They were caught while planning a bank robbery. As they interrogate them, the police keep them cut off from each other. Each of them is offered two options: to remain silent or to accuse their accomplice. They are told the following: if none of them speak, they shall both be free in a year; if only one of them do, she shall be freed immediately, while the other will spend ten years in jail; finally, if they accuse each other, they will be imprisoned for five years.

Scenario 3 is an instance of the famous Prisoner's Dilemma – one of the most discussed and analyzed applications of Game Theory. It is quite standard to label the two options offered to the participants "defect" (the option of accusing the accomplice) and, respectively, "cooperate" (the option to remain silent). And this seems intuitively all right. Is it what the proposed account predicts?

Presumably, if Rob decides to remain silent, this is because he believes and hopes that Octavia too will keep her mouth shut.[61] That is, if he remains silent, he pursues a goal (that none of them say a word) which he expects to share with Octavia. And, *mutatis mutandis*, the same is true of Octavia. Thus, if they both refrain from accusing each other, they both pursue a goal they have (and expect to have) in common, by following a plan they share (and think they share), a plan which has it that.... each should remain silent (forgive the repetition). That is, according to *View 3*, they cooperate.

On the other hand, if Octavia decides to accuse Rob and vice-versa, *View 3* correctly predicts that they do not cooperate. For, in such a situation, they fail to pursue a common goal. Here come what I take to be the less implausible candidates to be the shared goal that Octavia and Rob would allegedly pursue by playing "defect":

(1) *That they both accuse each other*: but Octavia and Rob do not share this goal, for none of them has it – none of them is interested in being accused by the other; none of them has any disposition to help bringing it about that the other choses the "defect" option, or to feel disappointment if the other doesn't play "defect", etc.[62]

(2) *That one of them accuses the other while the other remains silent*: neither Octavia nor Rob has this goal – having this goal requires being disposed to remain silent upon learning that the other has chosen the "defect" option and, clearly, none of them has this disposition.

[61] Here some may object that Rob might remain silent because accusing Octavia would be contrary to his code of honor; or because he wants Octavia out of jail more than himself. That might well be the case: but then the scenario wouldn't be an instance of the Prisoner's Dilemma. If Scenario 3 is to instantiate this pattern, we must assume that the agents involved are exclusively motivated by the years they would themselves spend in prison, given the different possible outcomes.

[62] Here one might object that I unduly assume that goals are distributive in the following sense: if X has the goal that A&B, then X has the goal that A and the goal that B. Albeit principles like this are rarely fully general, I find this one particularly difficult to prove wrong. This said, even if the principle isn't fully general, and even if it fails precisely in some situations corresponding to the Prisoner's Dilemma, I'm still on the safe side: my account not only requires that those who cooperate share a goal, but also that they share a plan. This, in turn, implies that each participant wants, hopes and believes that the others will play their parts. Applied to a Prisoner's Dilemma with a [defect, defect] outcome, this means that both player wants that the other plays "defect". And this isn't the case.

Thus it would seem that, if they chose to "defect", Rob and Octavia do not pursue a common goal. By consequent, they do not cooperate according to *View 3*.

Scenario 4:

Barak and Angela are walking down the street. At some point, they are on the same sidewalk, five meters away from each other, heading in the same direction, for about 500 hundred meters. They kind of notice each other, but they don't bother to talk to one another. After a while, Angela turns right while Barak keeps going straight.

Obviously enough, Barak and Angela do not cooperate in Scenario 4. They would cooperate if they were going for a walk together: but they aren't. The account under consideration supports and explains such a verdict: the protagonists of Scenario 4 don't share either a goal, or a plan. There is therefore no cooperation going on between them.

2.4. Objections and further developments

I shall now consider several objections to my proposal. Discussing them will, I hope, shed new light on the account.

2.4.1. Cooperation and strategic behavior

According to the literature (see (Andersson, 2007), (Petersson, 2007), (Bradsley, 2006), (Gold & Sugden, 2007), etc.) a major challenge for accounts of cooperation is to be able to distinguish mere strategic interactions, or interdependent individual actions, from genuine instances of cooperation.[63] Thus, it is natural to ask how my account meets this demand. At a general level, the account provides with a clean answer: in cooperation, but not in mere strategic interaction, the agents involved share both a goal and a plan. Or, to put it differently, whereas in mere strategic interaction participants

[63] To be sure: this doesn't presuppose that cooperation isn't a special case of strategic interaction (nor does it presuppose the contrary), but only that not all cases of strategic interaction are cooperative. I shall use "mere strategic interactions" to refer to the non-cooperative strategic interactions.

consider each other as independent agents who each pursue their own goals by their own means, cooperators intentionally pursue a single goal by means they agree upon. This sounds quite uncontroversial to me.

But, no matter how satisfactory the general answer I give rings, one might still believe that the details of my account are mistaken, i.e. that the particular rendering of shared goals and shared plans I've built into my proposal is mistaken and that, because of this, it classifies certain mere strategic interactions as instances of cooperation.[64]I consider several examples that some may deem troublesome.[65]

> *Oil or Solar Energy*:[66] Raeghar and Asha are competitors: they build cars and sell them on the European market. They face the following choice: they should either go for oil-based or for solar energy-based production (they won't be able to change afterwards, at least for some quite long time). For the time being, oil is cheaper; but they both know there isn't much left (in the whole world): thus, if they both decide to use oil, they are likely to go bankrupt. Most than anything, they want to avoid that. They also know that if one of them uses oil and the other uses solar energy, the former shall inevitably become the biggest car producer on the market. Finally, they both consider that – provided the other uses solar energy too – it would be all right to choose

[64] A first challenge from *mere strategic interaction* is to explain why, in a Prisoner's Dilemma situation, a "defect"-"defect" outcome isn't cooperative (see (Gold & Sugden, 2007, pp. 112–113)). I've done this in the previous section.

[65] My choice of examples is guided by game-theoretical considerations. In particular, I'll discuss one scenario corresponding to the Hawk-Dove game and several corresponding to the Battle of the sexes. As mentioned in the precedent footnote, I've discussed the Prisoner's Dilemma in Section 3.

[66] This scenario is an instance of the Hawk-Dove game:

		Player 2	
		Hawk	Dove
Player 1	Hawk	-5,-5	3,0
	Dove	0,3	2,2

Hawk-Dove has two Nash-equilibria (a Nash-equilibirum is an action profile, or a game outcome, such that none of the players can improve on her situation by unilaterally changing her play): [hawk, dove] and [dove, hawk]. (Gold & Sugden, 2007) argues that many influential accounts of collective behavior (mainly Tuomela's and Bratman's) wrongly reckons as collective behavior any outcome corresponding to such a Nash equilibrium. Here, I argue that my account of cooperation, a collective behavior if any, avoid this predicament.

the environment friendly option. Summing up: ideally, Raeghar would want that he uses oil and Asha uses solar energy; still, it would be kind of okay for him to use solar energy, provided Asha does it too; finally, if he knew that Asha will use oil, he would reluctantly go for solar energy: that they both exploit oil would be too much of a disaster. *Mutatis mutandis*, the same is true of Asha. Raeghar decides to make a move: he publicly announces he'll opt for an oil-based industry. He is half bluffing, but it works: Asha decides to go for solar energy.

Arguably, *Oil and Solar Energy* does not depict Raeghar and Asha as cooperating but as strategically interacting, in a rather antagonistic perspective. Thus, this scenario would indeed constitute a counterexample to the account put forward in Section 2.3, if it met the conditions set out in the proposal. But does it?

If someone claims that my account wrongly reckons *Oil or Solar Energy* as a case of cooperation, she must start by pointing out at least one goal that Raeghar and Asha share and pursue by acting the way they do. There are several plausible enough candidates: some may claim that they both pursue the goal that one of them goes for oil, while the other goes for solar energy (for both of them are somehow disposed to go for the option that they believe will be avoided by the other player);[67] or the goal that they do not both go for oil (for both of them are disposed to avoid the oil option, if they believe that it will be chosen by the other player); etc… I shall thus grant that *Oil or Solar Energy* meets the common goal requirement set up by my account.

But, beside the demand for a common goal, my account also requires that cooperators share a plan. Thus, my opponent also need to suggest a plan that could plausibly be the one that Raeghar and Asha share and follow towards the goal they have in common. And I don't believe she can do this. The reason why is simple: in *Oil or Solar Energy*, Raeghar ends up going for oil, while Asha choses solar energy. Thus, the shared plan they follow by so acting (no matter what the plan's goal is), if any, specifies that Raeghar should go for oil and Asha for solar energy. Hence, if Raeghar and Asha shared

[67] This goal corresponds to the two Nash-Equilibria of the Hawk-Dove game.

a plan, Asha would want Raeghar to go for oil.[68] And she doesn't: rather she would be better off if he went for solar energy (if there is any doubt here, have a look at the pay-off matrix in footnote 66).[69]My verdict is thus that *Oil or Solar Energy* is no threat to my account.

Let us now consider a different story:[70]

> *Red or Green:* Tomorrow is New Years' Eve. Ronaldo wants to wear red trousers for the party. Unfortunately, he thinks that Joshua might wear red trousers himself and he would rather not wear clothe the same color as Joshua's. Were he not to wear red pants, he would go for green trousers. Mutatis mutandis, Joshua is in the same state of mind. All of this is common knowledge among them. Joshua has an idea: he tells Alice, who is known to be a bigmouth, that he will definitely wear red trousers. Upon hearing this, Ronaldo decides to wear green trousers. Joshua hears of Ronaldo's decision, and hence settles on a red outfit for the party.[71]

Now, consider Ronaldo's and Joshua's states of mind once they've decided which trousers to wear on New Year's Eve's party. Ronaldo has decided to wear green trousers because he believes that Joshua's will be red. By acting this way, he pursues (among other things) the goal that he and Joshua do not have trousers the same color for the party. He thus desires and hopes that Joshua will not show up with green trousers (to his surprise). As for Joshua, he has decided to wear red trousers. By acting this way, he pursues (among other things) the goal that he and Ronaldo do not have trousers the same

[68] For, as per the definition of Section 3.2, if they share a plan, they both accept it; and if Asha accepts a plan which says that she should go for solar energy and Raeghar should go for oil, she wants, hopes or desires that Raeghar goes for oil.

[69] Importantly, this verdict seems to be independent of the specifics of the situation. Thus, it is to be expected that Nash-Equilibria outcomes in circumstances adequately modelled by the Hawk-Dove game never involve cooperation.

[70] This story instantiates the game known as the Battle of the Sexes. A matrix corresponding to this game is:

		Player 2	
		A	B
Player 1	A	3,1	0,0
	B	0,0	1,3

[71] Thanks to Olle Blomberg, Teresa Marques, Ljubomir Stevanovic, Esa Díaz-León, Dan Zeman, Max Kölbel and Dan López de Sa, for an insightful Facebook discussion.

color for the party. He believes and hopes that Ronaldo's trousers will be green. Thus, in *Red or Green*, Ronaldo and Joshua pursue a goal they share: that they do not wear trousers the same color for New Year's Eve's party. Furthermore, they pursue it by following a plan that they also share: Ronaldo should wear green pants, whereas Joshua should wear red pants. And they are well aware of all this. All the conditions set out by my account are thus satisfied. And yet, some will claim that, intuitively, no cooperation takes place in *Red or Green*. Is it then that *View 3* misses something? I don't think so. That is, although this isn't obvious, I think that Ronaldo and Joshua do cooperate in *Red or Green*. And I think that *View 3* suggests a satisfactory explanation of the fact that this verdict isn't very much intuitive. Let me explain.

View 3 aims at revealing the nature of cooperation. Thus, if the view is correct, we should expect the intuition that there is cooperation to co-vary with the intuition that the conditions put forward by the account are satisfied. In particular, *View 3* is perfectly compatible[72] with the fact that, when the elements of its analysans are both instantiated in a non-paradigmatic way and made hardly noticeable by other more salient features of the situation, intuition tends to lean toward a non-cooperation verdict. And this is what happens in *Red or Green*: neither is it intuitively clear that Joshua and Ronaldo cooperate, nor is it intuitively clear that they meet the conditions put forward by *View 3*.[73]

The fact is that, in *Red or Green*, although the protagonists share a goal they also have conflicting desires. Furthermore, several features of the scenario make their antagonism more salient than their common interest. Firstly, their shared goal isn't paradigmatic; rather, it is what we could call 'a shared goal by omission': it doesn't require that Joshua and Ronaldo act in a way that yield a certain outcome, but rather that they refrain from acting in a way that bring about a given state of affairs (hence, their shared goal would be automatically satisfied if either of them suddenly stopped existing). I'm not quite sure that this distinction is theoretically robust – but it is presumably significant when it comes to prompting an intuition or another. Secondly the link between their shared plan and the goal they share

[72] I would even say "predicts".
[73] A suggestion (made repeatedly) by Dan López de Sa is the origin of this reply.

is lose, at least on Joshua's end. That is, hadn't he had the goal he shares with Ronaldo, Joshua would nevertheless had performed the action the shared plan prescribes (i.e. putting on red trousers). Thirdly, Joshua's strategy is unfriendly. On the one hand, it involves providing incorrect information to Ronaldo. But, more importantly, it is unfriendly because it is chosen over a friendly and very natural option there is in the vicinity: talking to Ronaldo. Finally, the phrasing emphasizes the conflicting interests: the desire that grounds the conflicting interests is presented first, unhedged; the basis for the common goal is introduced next, nuanced by a 'rather'. And this suggests that Joshua cares more about wearing red trousers, than he cares about the goal he shares with Ronaldo.

If these features explain why *Red or Green* doesn't prompt the intuition that Joshua and Ronaldo cooperate (at least not to many people), we should expect the intuitions to vary if we modify them. And, I believe, the prediction is correct. Let us start by changing the phrasing only.

> *Red or Green – new phrasing:* Tomorrow is New Year's Eve. Both Ronaldo and Joshua are going to the party– and they know that. Joshua really doesn't want to wear trousers the same color as Ronaldo. He wants to wear either red or green pants, and to be honest, he would prefer red – but he thinks that Ronaldo may put on red pants himself. Mutatis mutandis, Ronaldo is in the same state of mind. All of this is common ground between them. Joshua has an idea: he tells Alice, who is known to be a bigmouth, that he will definitely wears red trousers. As expected, Alice ends up telling Ronaldo what Joshua told her, and Ronaldo thus decides to wear green trousers. Upon hearing this, Ronaldo decides to wear green trousers. Joshua hears of Ronaldo's decision, and hence settles on a red outfit for the party.

My appraisal of *Red or Green – new phrasing* is that, just by changing the way the story is told, we manage to suggest that Joshua's maneuver is intended as a solution to a coordination problem, rather than as a way to satisfy the desire which antagonizes him and Ronaldo. Accordingly, it doesn't seem that odd anymore to say that they cooperate. We can go a step further by making Joshua's move looks less unfriendly.

Red or Green – the friendly one: Tomorrow is New Year's Eve. Both Ronaldo and Joshua are going to the party – and they know that. Joshua really doesn't want to wear trousers the same color as Ronaldo. He wants to wear either red or green pants, and to be honest, he would prefer red – but he thinks that Ronaldo may put on red pants himself. Mutatis mutandis, Ronaldo is in the same state of mind. All of this is common ground between them. Unfortunately, they cannot get in touch directly with one another: they were in a relationship *together* for years and their new partners are tremendously jealous. Joshua has an idea: he tells Alice, who is known to be a bigmouth, that he will definitely wears red trousers. As expected, Alice ends up telling Ronaldo what Joshua told her, and Ronaldo thus decides to wear green trousers. Upon hearing this, Ronaldo decides to wear green trousers. Joshua hears of Ronaldo's decision, and hence settles on a red outfit for the party.

In *Red or Green – the friendly one*, Joshua's maneuver doesn't look as unfriendly as in the previous scenario. Importantly though, I haven't achieved this by changing the 'intrinsic nature' of the actions he performs in order to get his behavior and Ronaldo to be coordinated (in particular, he is still indirectly providing Ronaldo with incorrect information). Rather, I've made it explicit why he doesn't go for the natural, friendlier option (talking to Ronaldo). Hence, the maneuver appears to be less motivated by the antagonism, which in turn becomes less salient. Accordingly, it is even more intuitively plausible to say that Joshua and Ronaldo cooperate. To conclude, I shall make the goal and the plan Ronaldo and Joshua share stronger:

Red or Green – clearly cooperative: Tomorrow is New Year's Eve. Jana tells both Joshua and Ronaldo the following: "if one of you wear red trousers and the other wear green trousers, I'll offer you a whole week of holidays in Guadalupe. Of course, though, you cannot communicate in order to succeed." Both Joshua and Ronaldo don't have much money, and they both really want to go to Guadalupe. They wouldn't wear such a colorful outfit of their own initiative, but if it's a choice between red and green, they would both prefer their trousers to be red. All of

this is common knowledge between them. Joshua has an idea: he tells Alice, who is known to be a bigmouth, that he will definitely wear red trousers. As expected, Alice ends up telling Ronaldo what Joshua told her, and Ronaldo thus decides to wear green trousers. Joshua hears of Ronaldo's decision, and hence settles on a red outfit for the party.

As predicted, in *Red or Green – clearly cooperative*,[74] it is uncontroversial that Joshua and Ronaldo cooperate in order to reach the goal that they do not wear trousers the same color (and that one of them wears green pants, and the other red ones).

This reply of mine raises a question: why not enrich the account of cooperation and require (i) the absence of conflict, (ii) the presence of a paradigmatic shared goal, (iii) that the cooperators do not reach coordination thanks to 'unfriendly' maneuvers. The answer is, I believe, quite straightforward. Regarding (i) and (iii): real life cooperation almost always involves people with conflicting interests which negotiate in a not completely friendly way to decide how they are to pursue the goals they have in common. Regarding (ii): they are intuitively clear cases of cooperation grounded in shared goal 'by omission' (that is, the presence of a shared goal 'by omission' contributes but doesn't suffice to weaken the intuition that cooperation takes place). Furthermore, we don't need to do so: as pointed out earlier, if *View 3* is correct, it is to be expected that cooperation doesn't *obviously* take place, when the conditions the account puts forward are satisfied in a non-paradigmatic and non-salient way. Thus, theoretical simplicity recommends that we do not enrich the proposal any further.

2.4.2. Cooperation without a shared goal

According to the account of cooperation defended in Section 2.3, the agents who participate in a cooperative venture share its goal, i.e. each of them wants the goal to be achieved. In this sense, the view I put forward is similar to Bratman's account of shared action. Now, it has been forcefully argued

[74] If you wonder what this scenario still has in common with the original Red or Green, notice that they both correspond to the same game-theoretical situation (the Battle of the Sexes whose matrix you will find in footnote 70). Furthermore, in both cases, the situation resolves into the same Nash Equilibrium: Joshua "plays" red while Ronaldo "plays" green.

(see (Kutz, 2000) and (Shapiro, 2014)) that Bratman's proposal is inadequate to account for shared activities in hierarchical contexts. In such situations, in which some people follow orders rather than their own inclinations, agents are likely to end up participating in collective projects whose goal they do not share – or so it has been suggested. Thus, given that the phenomenon Bratman analyzes is a close relative of the phenomenon I target, it is natural to wonder whether an analogous objection would affect my account of cooperation.

Consider an example of scenario that could be taken to motivate the objection:

> *A coup*: As part of their plan to overthrow the government, some people offer me one million euros to hack the prime minister's e-mail account. One million euros seem worth the risk and I accept. I know what the conspirers' goal is, but I'm not sympathetic to it: I like the government, and I find the conspirers chaotic – I believe they would be very poor governors. To be honest, I would want them to fail and I even feel a bit guilty about accepting their offer – but it is such a lot of money!

According to my opponent, in *A coup* I do cooperate in the conspiracy by hacking the minister's account; and yet, I do not share the conspiracy's goal, for I would rather want the coup to fail.

I reject this reading of *A coup*. I believe that, as stated, the scenario is underdetermined in respects crucial to the present issue. It leaves open whether I want or fail to want the coup to be successful. And it doesn't settle whether I'm cooperating in the coup or not. Once we focus on fixing these features, we see that there is no pressure to accept that I could cooperate in the coup without *somehow* desiring it to be successful.

Admittedly, it may seem strange to claim that, as stated, *A coup* is underdetermined with respect to whether I want or fail to want the conspiracy to be successful. After all, it is said explicitly that I'm not sympathetic to the conspirer's goal and that I would rather want them to fail. But closer scrutiny reveals that this first impression is mistaken. We, humans, quite normally host conflicting desires: that I wish to be a vegetarian doesn't mean that I don't also feel like eating a hamburger; that I desire to

wake up early tomorrow and seize the day doesn't mean that I don't also feel like sleeping in tomorrow morning; etc. Similarly, that I would want the conspirers to fail doesn't mean that I don't also want them to succeed. Thus, there are ways of making *A coup* more determinate (henceforth "determinations") according to which I do somehow desire the success of the conspiracy, and others according to which I do not.

Now, there are determinations of *A coup* in which I both desire the conspiracy to be successful and cooperate in it. Say for instance that, after accepting the deal, I start fearing the reprisals of the current government, in case the conspirers miss their shot. Moved by this fear I help the conspirers as much as I can, besides hacking the minister's account: I tell them everything I learnt about the government informatics system, I explain to them in which circumstances could the hacking be discovered, and how they (the conspirers) could check whether it has been discovered, etc. In such situation, even though I believe that the conspirers would be poor governors whereas the current government is quite alright, I cooperate in the coup.

There are also determinations of *A coup* in which I do not want the conspirers to succeed and I do not cooperate with them, in spite of performing my part of their plan to overthrow the government. In the most obvious example, I know that, by hacking the prime minister's e-mail account, I will trigger a security procedure which will almost certainly wind up in the failure of the conspiracy and the imprisonment of the conspirers. In this case, I don't desire the conspirers' success (or at the very least, I do not act on this desire) and I do not cooperate in the conspiracy, *in spite of doing what the conspirers pay me for*.

But, on the other hand, there are no determinations of *A coup* in which I do not want in any sense the conspirers to succeed and yet I cooperate with them. For I don't cooperate in the coup if I don't have some kind of predisposition to do more than strictly and unreflectingly performing the task that has been allocated to me; and that I have such predisposition means precisely that, somehow, I desire the success of the plan I'm following.

My contention is that all the putative counterexamples to the claim that agents must somehow desire the success of the projects they cooperate in will draw on the same kind of indeterminacy diagnosed in *A coup*. The scenario will make it explicit that, among those agents who cooperate in a given project, there is someone who, *in one sense*, desires the project to be a failure. From this, it will be inferred that this cooperator doesn't desire the project to be successful *in any sense*, thus offering a counterexample to the account I uphold. But reflection on the nature of desire shall reveal the invalidity of the inference: in some determinations, the desire for failure will be accompanied by a conflicting desire for success, and these shall be the determinations in which the agent under scrutiny can be genuinely described as cooperating.

2.4.3. Spontaneous cooperation: no need for a plan

Some will think that spontaneous cooperation poses a special problem to my account; or more explicitly that, as all "planning" accounts of cooperation, mine fails when it comes to analyzing spontaneous (i.e. unplanned, or so it would seem) instances of cooperation. For instance, when I meet my friend Emiliano, we systematically high-five and, although this is clearly a cooperative activity of ours, it is dubious that we plan it in advance. Or, consider a situation in which two people rush to help a third person who is having trouble getting her wheelchair on a bus (see (Blomberg, 2013, p. 28-31)): again, it would seem that they are cooperating, but without any planning taking place prior to their actions.

This objection presupposes that, when someone acts on a plan she accepts, she must accept that plan before her action takes place. Unsurprisingly, I'll reject this assumption. If my brother sends a ball in my direction shouting, "Catch it!", and if I start running towards it because I want to do so, I'm hereby accepting and following a plan – which has it that, in order to catch the ball, I should run *this* way (I might not be able to give a non-indexical description of the instructions of the plan). In the same sense, upon seeing my friend Emiliano, I form the desire that we high-five, determine that high-fiving requires that we raise our hands in such-and-such way (again, I may not be capable of individuating such a way

non-indexically) while being responsive to each other's movements, and thus decide to raise my hand in the appropriate way, hoping that Emiliano will follow. When Emiliano sees me, he realizes which plan I'm following and opts for going along. When this happens, we start cooperating. And I expect that a similar account will be available in analogous cases of 'spontaneous' cooperation. I grant that there is something odd in saying that, even in spontaneous cooperative episodes, participants follow a plan. But I think the oddity just suggests that the notion of plan I work with is more inclusive than the ordinary notion.

2.4.4. Animals and children: the threat of intellectualism

An important question that any account of cooperation must be confronted with is whether it can accommodate the intuition that cooperation often takes place among children and animals or – more generally – agents who aren't in possession of what might be called a "robust theory of mind", i.e. a sophisticated conceptual scheme to represent other agents as endowed with a mental life. The issue is best introduced by quoting Tollefsen, who has forcefully highlighted the limitations of accounts of joint intentions[75] that impose too demanding cognitive requirements.

> "There are a variety of accounts of joint intention on offer. Most of these accounts analyze joint intention in terms of a highly complex set of individual intentional states that exist under conditions of common knowledge. In this article, I want to pose a problem for these accounts … I want to argue that these accounts are too complex to accommodate cooperative and joint actions when participants are animals and young children (ages 1 to 4). In particular, I will argue that these accounts presuppose that the participants in a joint action have a robust theory of mind. Children under the age of 4 (possibly 5) and animals lack a robust theory of mind… But they do engage in joint action, or so I shall argue. This suggests that the requirements for joint intentional action need to be weakened and revised." (Tollefsen, 2005, p. 76)

Tollefsen's argument is crystal clear. Firstly, she claims that animals and children under the age of 4 engage in cooperative activities. Secondly, she argues that both animals and young children lack a robust theory of mind. Finally, she concludes that any theory of cooperation that requires that the participants be in possession of a robust theory of mind at best states sufficient conditions and thus, that, if we want to have a general account, we need to "weaken and revise" such theories.

[75] Roughly, a joint intention is the network of intentional states who drive agents to cooperate.

I'm willing to grant Tollefsen's first premise. It is prima facie very tempting to say that hyenas and wolves engage in cooperative hunting, for instance; and some children under the age of 4 appear to readily communicate, a paradigm of cooperative activity. Such intuitive judgments can always be rejected on theoretical grounds, but it is beyond reasonable doubt that it would be an important shortcoming of my account of cooperation, if it were incapable of accommodating them. As for the second premise, I'm unfortunately in no position to assess it. My knowledge of experimental cognitive sciences is miles away from enabling an informed verdict. Thus, my strategy will be to take for granted the claim that children and non-human mammals lack a robust theory of mind, and try and argue that the framework set up by my account of cooperation can nevertheless accommodate the intuition that they do cooperate.

According to Tollefsen,

> "the central features of a robust theory of mind (..) include the following:
> 1. an understanding of other persons in terms of their thoughts, intentions, and beliefs;
> 2. an understanding that other persons' thoughts, beliefs, and intentions may differ from one's own; and
> 3. an understanding that others have thoughts and beliefs that may not match with the current state of affairs (false beliefs)." (Tollefsen, 2005, p. 81)

Albeit Tollefsen doesn't make this explicit, it is natural to assume that she considers these conditions as both sufficient and necessary. Thus, it would seem that my account avoids the blade of Tollefsen's argument if it doesn't require that cooperators satisfy conditions 1 to 3 above. To ease the discussion, let me rehearse the account once again:

> COOPERATION: The Xs cooperate if and only if, for some J, (A) the Xs share the goal that they bring about J, (B) the Xs share a plan for them to J, (C) each X pursues the Xs' goal, by following the Xs' plan, and (D) it is overtly believed among the Xs that (A)-(C) hold.

Clearly enough, satisfying conditions (A) to (D) is only possible if one possesses *some kind* of a theory of mind. On the one hand, the account clearly requires that each cooperator conceives of the other participants as capable of *doing things*, i.e. as agents – at least in some rudimentary sense. This is so because, for instance, clause (A) demands that each cooperator wants all the cooperators to perform

certain actions (actions leading to their J-ing). And, when we want something to perform an action, we presumably represent that thing as an agent. On the other hand, clause (D) requires that each cooperator have beliefs regarding, firstly, the goals of the other participants; secondly, the way in which they hope to achieve these goals; and, thirdly, some beliefs they hold about each other's goals and plans. This is, I must admit, an impressive list: in particular, it appears to entail that each cooperator has an "understanding of other persons in terms of their thoughts, intentions and beliefs"(Tollefsen, 2005, p. 81). Still, I believe that our initial verdict must be that my account does not require a robust theory of mind for at no point does it demand that the cooperators be capable of representing other agents as having beliefs and intentions different from their owns, or beliefs which contradict the actual state of affairs. Quite the opposite indeed: (clause (D) of) my account demands that the cooperators represent each other as having the same goal, the same plan to achieve it, and the same iterated beliefs.

Unsurprisingly though, things aren't that simple and Tollefsen's argument cannot be dismissed so quickly. For many people will claim (and Tollefsen seems to be among them) that an agent can only represent someone as believing or intending if she can represent someone as having intentions different from hers and beliefs which contradict the beliefs she holds or the actual state of affairs. Indeed, experiments designed to determine whether children of a certain age are capable or not of representing others as *believing* or *desiring* appear to systematically test the capacity to attribute beliefs and goals different from one's own (see, for instance, (Gopnik & Slaughter, 1991) and (Wellman & Bartsch, 1988)). And, crucially, Tollefsen claims that children start engaging in cooperative activities at an age (under 4) at which they are still incapable of attributing either false beliefs, or beliefs different from the ones they hold. We thus have the form of an argument against the account of cooperation developed in Section 2.3:

> *Premise 1:* An agent represents other agents as having beliefs only if she can attribute both false beliefs and beliefs different from her own to other agents.

Premise 2: According to COOPERATION, cooperators represent each other as having certain beliefs.

Intermediary conclusion from P1 and P2: According to COOPERATION, cooperators must be capable of representing other agents as having both false beliefs and beliefs different from their own.

Premise 3: Children begin to engage in cooperative activities before they are capable of representing other agents as having either false beliefs or beliefs different from their own.

Conclusion: COOPERATION isn't fully general since it cannot account for cooperation among young children.

There are, I take it, two strategies open to me in reply to this argument. Firstly, I may challenge Premise 1 along the following lines: granted, understanding what mental representations (i.e. beliefs) are consists in part in understanding that they may differ from reality (and from one's own). Granted, representing someone as believing requires that one somehow understands the nature of beliefs. But it only follows from this that representing someone as believing entails that one somehow understands that beliefs can be false, not that one can actually attribute false beliefs, or beliefs conflicting with one's owns. Arguably, the latter capacity may involve other cognitive resources, whose absence or insufficiency may well account for failure in the false-belief task and related tests (for an argument in this direction, see (Bloom & German, 2000)). But although it is definitely an interesting reply, it crucially hinges on what is the correct interpretation of certain empirical findings (the results of the false belief task and related tests). In particular, it does challenge the claim – which I promised to leave untouched – that children under the age of 4 (and animals) do not have a robust theory of mind. For this reason, I'm not willing to rely on it too heavily.

The second strategy is, I take it, more promising. It argues that, according to a very natural reading of COOPERATION, while *full-blown* cooperation requires a *full-blown* theory of mind, rudimentary cooperation shall only require a rudimentary theory of mind. In other words: the reply insists that

COOPERATION highlights some conceptual connections between the notion of cooperation and other notions (shared goal, shared plan and overt belief); thus, the reply continues, when cooperation is applied in a less demanding or looser sense, it is to be expected that the notions we used in its analysis only apply in a looser sense too.

For this strategy to be effective, two conditions must obtain. Firstly, it should be plausible that when we say that great mammals and children under 4 cooperate, we use the notion in a somehow undemanding sense. Secondly, it should be plausible that great mammals and children can meet the conditions established in COOPERATION, provided that they are understood in a correspondingly undemanding sense.

I haven't much to say to support the claim that, when applied to very young children and higher mammals, the notion of cooperation is used somehow loosely. It strikes me as very plausible: as I've argued at some length in Section 2.3, I take it that cooperation intuitively involves a sense of sharedness, which in turn requires an understanding of other agents' minds. Hence, insofar as we apply this notion to agents who only have a very rudimentary sense of sharedness (if any), my understanding is that we use the notion in a rudimentary or undemanding way.

As for the second claim, I believe that there are plenty of reasons to think that young children possess a rudimentary theory of mind which makes it possible for them to satisfy the conditions COOPERATION puts forward, provided that there are understood in a relaxed way. Hence, children under the age of 4 who still fail the false-belief test and related task already employ the vocabulary we use to refer to belief states (e.g. they say things like "he thinks that BLA").[76] Admittedly, they aren't completely proficient with these terms – they make mistakes in a systematic and predictable way. But their usage is far from being completely random, and it thus seems right to say that it reveals at least some understanding of the notions the terms refer to. Furthermore, children under the age of 4 seem quite proficient in using other notions (such as goals, intentions and desires) which are, presumably,

[76] If you have any doubt, check this out: https://www.youtube.com/watch?v=8hLubgpY2_w.

conceptually connected with the notion of beliefs. In particular, and as Tollefsen underwrites (Tollefsen, 2005, p. 87), children under the age of 4 are capable of recognizing the intentions with which people perform certain actions. And this, *pace* Tollefsen, seems hard to square with the claim that they do not also somehow conceive of other agents as believers. Consider a toy example: Martin, 3 years old, looks at Roberto, 44 years old, who is trying to break a window with a big stick; Martin recognizes Roberto's intention, i.e. he understands what Martin is trying to do. Does it make sense to claim that, when he thinks of Roberto as trying to break the window by hitting it with a big stick, Martin isn't also somehow thinking of Roberto as believing that he can break the window by so acting? I don't think so. In general terms, I believe it is a conceptual matter of fact that, by representing someone as trying to achieve G in a way W, I *somehow* represents her as believing that W is a way to achieve G.[77] Finally, by the age at which they start engaging in cooperative activities, children appear to be capable of engaging in *joint attention*[78], where joint attention requires that "two individuals know that they are attending to something in common" (see (Tomassello, 1995, p. 106). Furthermore, Tollefsen argues that, in order to "introduce the openness that needs to be present in cases of joint action" (Tollefsen, 2005, p. 92) we should follow Peacocke (Peacocke, 2005) and characterize joint attention as follows:

> "X perceives that x and y are attending to o.
> Y perceives that x and y are attending to o.
> X perceives that y perceives that x and y are attending to o.
> Y perceives that x perceives that x and y are attending to o.
> X perceives that y perceives that x perceives that x and y are attending to
> o . . . and so on." (Tollefsen, 2005, p. 92)

[77] Tollefsen is well aware that she puts forward empirical claims that some will find hardly compatible (namely the claims that, under the age of 4, children recognize intentions whereas they are still incapable of attributing beliefs). Nevertheless, she insists that "there is strong experimental evidence that suggests that young children have an understanding of others' intentions and desires much earlier than they acquire an understanding of belief" (Tollefsen, 2005, p. 88) and she therefore ventures that it may be wise to abandon the thought that intentions and beliefs are conceptually intertwined. But, I take it, the empirical data that children under the age of 4 fare poorly at attributing false beliefs and beliefs different from their own falls short of showing that they don't have some rough understanding of beliefs. Thus, they put no pressure on the conceptual connection between intentions and belief I'm appealing to. Furthermore, there seems to be empirical evidence that, precisely, young children who still fail the false belief test have a rudimentary understanding of the interplay between beliefs and intentions (see (Moses, 1993)).

[78] At any rate, this is something that Tollefsen takes for granted (Tollefsen, 2005, pp. 24–25) and, remember, I'm taking Tollefsen's empirical claims for granted here.

And, as far as I can tell, being capable of engaging in joint attention so understood is being capable of participating in a rudimentary form of overt belief.

Admittedly, the matter may be more complex in the case of higher-mammals. There seems to be solid evidence that chimps have a rudimentary theory of mind, as forcefully claimed by Call and Tomassello:

> "even if chimpanzees do not understand false beliefs, they clearly do not just perceive the surface behavior of others and learn mindless behavioral rules as a result. All of the evidence reviewed here suggests that chimpanzees understand both the goals and intentions of others as well as the perception and knowledge of others." (Call & Tomasello, 2008, p. 191)

But we do not only say that chimps cooperate; we also use the idioms of cooperation when talking about hyenas, lionesses, etc.[79] and one may doubt that these mammals are sophisticated enough to read each other's minds, even in a rudimentary way. Furthermore, there seems to be evidence that full joint attention as characterized above isn't even present in apes (see (Carpenter & Call, 2013)), and thus that even apes may be unable of rudimentary satisfying the overt belief condition.[80]

Hence, in order to accommodate the intuitions that higher-mammals cooperate, I may have to stretch my rejoinder a bit further. Let me explain: I've said that, given the conceptual connections COOPERATION highlights, it is to be expected that a loose understanding of the notion of cooperation applies to situations in which the conditions COOPERATION puts forward are satisfied in a rudimentary way. I shall now take this reasoning a step further and observe that, by the light of the same conceptual connections, it is to be expected that, whenever it is intuitive or natural to describe a situation in terms of the conditions COOPERATION puts forward (irrespective of whether this description is *really* true), it shall be intuitive or natural to describe this situation as one in which cooperation takes place (irrespective of whether cooperation *really* takes place in this situation). Hence, since it is natural to describe the behavior of higher-mammals by appealing to the kind of intentional vocabulary

[79] Indeed, as pointed out in Chapter 1, we also use these idioms to talk about some insects (e.g. termites and ants).

[80] This being said, they may still satisfy a weaker notion of openness, one that Martin Davis (see (Davies, 1987)) has labelled "mutual absence of doubt", and which definition reads that: X and Y are in a state of mutual absence of doubt with respect to P if and only if (1) neither X nor Y doubt that P, (2) neither X nor Y think that either X or Y doubt that P, (3) ...

COOPERATION appeals to, this account doesn't clash with the intuition that higher-mammals cooperate, but rather predicts it (even though it may entail that the content of this intuition isn't *really* correct).

Chapter 3: Grounding institutional rules: power, commitment and publicity

---

Abstract: Institutional reality is that aspect of social reality which involves money, borders, legal and educational systems, driving codes, etc. Searle, Hindriks and Thomasson have extensively and convincingly, I believe, argued that institutional reality can be accounted for in terms of rules. Hence, for instance, their view implies that there is money in the United States (partly) because it is a rule of this country that some pieces of paper which meet certain conditions give certain rights and duties (characteristic of money) to their owners. Unfortunately, their proposals are wanting when it comes to making explicit what it takes for a given rule to be a rule of a country (or, for what is worth, a tribe, a hospital, or a village). Here, they appeal to the notion of collective acceptance, without offering elucidations thereof that would give it enough content to play the role it is meant to play. In this chapter, I fill in this loophole in the Searle-Hindriks-Thomasson approach to institutional reality by formulating and defending an account of institutional rules in terms of power, commitment and publicity and overt belief.

3.1. Introduction

Institutional reality is one of the most interesting and most studied aspects of social reality. It englobes money, borders, legal and educational systems, driving codes, etc. On a different level, it also contains the fact the Trump is the president of the U.S, the fact that a piece of paper I have now in my pocket is money, and the fact that Ruth Bader Ginsburg is a U.S. Supreme Court Justice.

In their respective work on the topic, Searle (see (Searle, 1995), (Searle, 2005) and (Searle, 2009)), Hindriks (see (Hindriks, 2009)) and Thomasson (see (Thomasson, 2009), (Thomasson, 2003a) and (Thomasson, 2003b)) have extensively argued that the core of institutional reality is constituted by collectively accepted or recognized sets of rules. Hence, their accounts have it that:

- There are U.S. Supreme Court Justices because (1) it is collectively accepted that whoever meets conditions **J** has a set of rights, obligations, duties, etc. (for short: deontic powers) which correspond to being a U.S. Supreme Court Justice and (2) some people meet conditions **J**.
- The piece of paper I have now in my pocket is money because (1) it is collectively accepted that whatever piece of paper which meets conditions **M** confers to her owner a set of deontic powers characteristic of money and (2) the piece of paper I have now in my pocket meets conditions **M**.
- Some actions are murders because (1) it is collectively accepted that actions meeting conditions **A** have normative consequences distinctive of murder and (2) some actions meet conditions **A**.
- Some events are elections because (1) it is collectively accepted that events meeting conditions **E** have normative consequences characteristic of elections and (2) some events meet conditions **E**.

In a schematic and very rough way, we can therefore sum up the Searle-Hindriks-Thomasson approach to institutional reality as follows:

1. People ground institutional reality by collectively accepting certain rules.
2. The rules collectively accepted by a given group of people (a) fix a set of institutional statuses (e.g. being a Supreme Court Justice, being money, being a murder, etc.), (b) specify the normative attributes linked to these statuses (e.g. having the power to settle disputes regarding the constitutionality of certain laws, giving the rights to her owner to buy items for a certain value, implying that her perpetrator is liable to a certain punishment), (c) determine the conditions in which these statuses are instantiated.

3. Institutional facts – facts about certain people, events, abstracts entities or objects instantiating institutional statuses – obtain because certain people, events, abstracts entities or objects meet the conditions set up by the collectively accepted rules.

I believe that the Searle-Hindriks-Thomasson approach to institutional reality is somehow on the right track. That is: I believe that rules are central to institutional reality in a sense that implies that, if we can account for the fact that a given society has the institutional rules it has, then we have a full account of that society's institutional reality. At any rate, I shall take this much for granted in this chapter.

I remain unsatisfied though with the claim that the institutional rules of a society S are those which are collectively accepted in S. For, as far as I see it, the pre-theoretical content of the phrase "collective acceptance" is too poor to offer anything worth calling an account. Hence, this claim calls for a theoretical elaboration of the notion of collective acceptance – which is unfortunately mostly missing in the work of Searle, Hindriks and Thomasson.[81]

My goal in this chapter is therefore to fill in this loophole in the Searle-Hindriks-Thomasson approach to institutional reality. That is, I shall try to answer the following question: what makes it the case that a given society has the institutional rules it has (rather than some other institutional rules or none)?

My plan is as follow: in Section 3.2, I briefly discuss the intuitive notion of institutional rules and introduce the idea of institutional anchor, which will help me frame my arguments. In Section 3.3, I submit some requirements and desiderata that any account of institutional rules should satisfy. Finally, in Section 3.4, I present and motivate an account of institutional rules in terms of power, commitment and publicity.

[81] To be fair, Searle twice offered some kind of elaboration of this notion. In his *The Construction of Social Reality* he appears to be claiming that collective acceptance must involve a primitive capacity of the human mind which he labels *collective intentionality* and which he sees as crucially at work in cooperation. But, as he himself recognizes in later work, this first stab is probably mistaken (see (Searle, 2009, pp. 56–58)). Then, in his *Making the Social World*, he says in passing that universal individual acceptance should suffice for collective acceptance (see Ibid.). This may well be true, but I take to be clear that, in the light of the discussion in Section 3.3., this cannot constitute a fully general account.

*3.2. Institutional rules as rules **of** or rules **adopted by** institutional anchors*

Many (if not any) social contexts, social groups and societies have rules. Hence, for instance, universities, hospitals, companies, etc. have regulations – and so do research groups, seminars, and so on. At a bigger scale, states have laws and there (presumably) are international laws which govern the interactions among them. Finally, at a smaller and less formal scale, kids playing in a schoolyard may develop quite determinate sets of rules – hence one of them may be the boss of the yard, and failure to show her proper respect severely punished. These are all examples of what I shall label *institutional rules*.[82]

It is important to notice from the outset of our investigation that what makes a rule institutional isn't an intrinsic feature of the rule itself, but rather something like it's being adopted (or collectively accepted, to remember the Searle-Hindriks-Thomasson terminology) by a given social group (or society). Consider, for instance, the rule which says that whoever possesses three red frogs has the right to free healthcare for the rest of her life. In the actual world, this is presumably no institutional rule (at least no earthly institutional rule). And that this is so isn't due to any intrinsic feature of this rule, but rather to the fact that no actual (earthly) social group has adopted this rule as one of its owns. On the contrary, the ban on smoking in public spaces actually is an institutional rule, because many states have adopted such a ban as one of their rules (e.g. Spain, Switzerland, etc.).

But saying that institutional rules are rules that are adopted by social groups (or societies) or, for short, rules of social groups (or societies) may be too quick. For many institutional rules aren't easily seen as rules of *social groups* (or *societies*). Consider for instance the regulation of a multinational company: which is the social group which adopts this regulation as its institutional rules? Is it the company's board? The company's employees? The company's clients? All of them? Or, for another similar example, think of the regulation of a private hospital: which is the social group which adopts this

[82] Admittedly, as used in ordinary parlance, the adjective "institutional" suggests a degree of codification and formalization that some of the examples above may fail to exhibit (e.g. the schoolyard case). Hence, my usage of the expression may require that we stretch a little bit the ordinary understanding.

regulation as its sets of institutional rules? Is it the hospital's board? The hospital staff? The hospital's clients? All of them? At any rate, in both these cases, the regulations in question are best described as the rules of, respectively, the company and the hospital. And it is at the very least controversial that the company and the hospital are identical to social groups (or for what matters, societies) in a way that would ensure the generality of the claim that institutional rules are rules *of* social groups (or *societies*).

Hence, while the claim that institutional rules are rules that are somehow adopted by something (or some things) as its rules (or their rules) seems correct, it also appears that we must be liberal when it comes to the kinds of entity which can have institutional rules. Thus, on the face of it:

- Social groups (groups of kids, groups of gangsters, small tribes, etc.) may have institutional rules
- Societies (the Spanish society, the American society) may have institutional rules
- Peoples (the Spanish people, the American people) may have institutional rules (hence, the Spanish Constitution presents itself as approved by the Spanish people)
- Entities such as companies, hospitals, universities and states may have institutional rules

To refer generically to the kind of entities which can have institutional rules, I suggest that we use the term "institutional anchor". Hence, the goal of this chapter is to answer the following question:

> Given an institutional anchor A, which has the institutional rules I_R, what makes it the case (or grounds) the fact that I_R are A's institutional rules?

At this stage of our inquiry, we don't know much about institutional anchors. In particular, we would be at pain to deliver the conditions something must meet to be an institutional anchor. Nevertheless, it seems obvious – and shall become relevant in the subsequent sections – that institutional anchors involve people. Hence, social groups, societies and peoples appear to be *made of* people; companies, hospitals, universities have employees and deliver services to clients; and states have citizens, legal

residents, etc. The general reason behind this seems to be that institutional anchors must be capable of adopting certain rules, and – intuitively – it takes people to perform this operation (or, at least, sophisticated enough agents).

3.3. Preliminary discussion: requirements and desiderata

As said earlier, the task I set out to myself in this chapter is to find out what makes it the case that, given an institutional anchor A with a set of institutional rules I_R, I_R are A's institutional rules. Or, in other terms, the goal I pursue here is to fill in the right-hand-side of the following biconditional:

> Given an institutional anchor A and a set of rules I_R, I_R are A's institutional rules if and only if….

Now, instead of starting directly by formulating an account, I shall start with a preliminary discussion which will narrow down the space of possibilities. Hence, I shall bring to the fore some obvious truths about the matter that interests us and discuss the requirements they impose on putative accounts of institutional facts. I shall also submit some (more controversial) desiderata that, I believe, a good account of institutional rules may be expected to meet.

3.3.1. Individual ignorance of institutional rules

Individuals (i.e. people) may ignore some (or most) of the rules of the institutional anchors they are involved in. Or, in more schematic terms, a rule R can be one of the institutional rules of an institutional anchor A even though many (or most, or all?) of the people involved in A don't know that.

Suppose, for instance, that the regulation of the *Universitat de Barcelona* (for short, UB) contains a rule stating that the teaching staff should change the password of their professional e-mail every three months. Suppose furthermore that most members of the UB's teaching staff didn't read the university's regulation. In such circumstances, a rule is in place in the UB, even though many people involved in this institutional anchor aren't aware of that.

Legal systems offer another obvious example. In most contemporary societies, legal systems are extremely complex – so complex indeed that, presumably, no single member of these societies is aware of all the rules they contain. And it is probably for this reason that such systems usually countenance a rule which states that ignorance of the law is no excuse or defense.

Thus, our account of institutional rules should make room for the possibility that, given an institutional anchor A and its rules I_R, many (or most) of the people involved in A don't know that I_R are A's institutional rules.[83]

3.3.2. Individual violation

Individuals can violate the rules of the institutional anchors they are involved in. Violation of institutional rules is indeed a very mundane phenomenon (think of how often we violate, for instance, the traffic codes of the country we live in), and most institutional systems have rules which specify how such violations are to be punished.

Furthermore, individual violations of institutional rules need not be a marginal phenomenon. Consider several examples:

> According to the tax system of State E, the E residents must declare all their incomes in order to pay taxes corresponding to the totality of their earnings. But most E residents only declare part of the money they are making in order to avoid paying all the taxes that are due.
>
> According to the tyrannical State G, none is allowed to criticize G's government in G's territory. But most of G's resident secretly and constantly criticize their government's actions.

[83] The fact that people can ignore the institutional rules of the anchor they are involved in is one of the reasons why Lewis' account of conventions cannot cover institutional rules. For Lewis's account requires that conventions be commonly known (see (Lewis, 1969, pp. 52–54)). Obviously, this is no objection to Lewis.

> In the city of Barcelona, it is forbidden to park one's motorcycle on a sidewalk that is less than three meters wide. But it is common knowledge among Barcelona's residents that the police rarely enforce this rule outside the city center. Hence, most people do park their motorcycle on such sidewalks as soon as they leave this area.

Hence, an account of institutional rules should allow for the possibility that at least some of the rules of an institutional anchor A be massively violated by the people involved in A.[84]

3.3.3. Individual dissatisfaction

Individuals may be fairly dissatisfied about the institutional rules of the institutional anchors they are involved in. They may find them absurd, unfair and/or inefficient and illegitimate. Consider yet another series of examples:

> The tyrannical State G forbids criticizing G's government. And yet, most of G's citizen find this prohibition horrendous and totally illegitimate.

> The slavery rules of State S require that slaves blindly obey their master. Slaves constitute ninety percent of S's population, and they find it deeply unfair that they have to obey their masters, whose authority they don't see as legitimate in any sense.

Hence, an account of institutional rules should allow for the possibility that most (if not all) the people involved in an institutional anchor A be deeply dissatisfied with A's institutional rules.

3.3.4. Unfairness

The rules of an institutional anchor may be completely unfair and illegitimate. Unfortunately enough, there are plenty of actual examples: the American Slavery Laws of the 19th century, the Nazi Laws in place during World War II, and many others.

[84] The fact that institutional rules may be massively violated is another reason why Lewis account of conventions, which identify the latter as behavioral regularities (see (Lewis, 1969, p. 78)), cannot cover institutional rules.

3.3.5. Selective enforcement or lack thereof

Institutional anchors may enforce their rules only selectively and it even appear that they may simply fail to enforce them. Consider:

> It is a rule of the traffic code of the Barcelona Municipality that motorcycles cannot be parked on sidewalks whose widths doesn't reach three meters. But it is also a well-known fact that this rule is almost never enforced outside the city center.
>
> By definition, jaywalking is forbidden by any traffic code. And yet – rare are the institutional anchors which enforce this prohibition.
>
> It is a rule of the Spanish Parliament that the representatives must address each other in Castellan. But, since Patxi López came to be the Parliament President, this rule isn't enforced anymore.

Hence, an account of institutional rule, should make room for the possibility that an institutional anchor enforces one (some) of its rules selectively, seldom, or even never.

3.3.6. Non-fundamentality

Institutional reality isn't fundamental, and the facts about certain institutional anchors having certain institutional rules aren't fundamental either. This claim may be fleshed out in many different ways. Some will say that institutional reality supervenes on other, more fundamental, aspects of reality; others will claim that institutional reality is determined by other more fundamental aspects or reality; and some will insist that institutional reality is nothing over and above other more fundamental aspects of reality. We need not worry about the details here (but, for a very detailed discussion, see (Correia & Schnieder, 2012)); the unspecific, uncontroversial and elusive thought that institutional reality isn't fundamental suffices for our purposes (indeed, this unspecific claim suits our purposes better than any of its specifications, since the latter shall invariantly be more controversial than the former).

A good account of institutional rules should reflect the fact that institutional reality isn't fundamental. In particular, such an account should suggest a story about the way how institutional reality emerges from (or is made of, constituted by, etc.) other, intuitively more fundamental, aspects of reality.

Because of the elusiveness of the notions involved, it isn't very clear which accounts would be ruled out by this desideratum (henceforth, the non-fundamentality desideratum), and which would fit the bill. I shall therefore discuss two examples, to try improving our understanding.

> Consider an account which reads that I_R are the institutional rules of an institutional anchor A if and only if most people involved in A individually follow I_R. This account meets the non-fundamentality desideratum – for it suggests that institutional reality emerges out of people who follow the same rules (unfortunately, this account is clearly false).
>
> Consider an account which reads that I_R are the institutional rules of an institutional anchor A if and only if some people who have the right to fix A's rules have chosen I_R to play this role. This account presumably fails to meet the non-fundamentality requirement: for, presumably, that some people have the right to fix A's rules is an institutional feature of reality. Hence, again presumably, the account fails to account for the existence of institutional reality in terms of more fundamental aspects of reality.

3.3.7. Avoiding Circularity

Consider the following putative accounts of institutional rules:

> Given an institutional anchor A and a set of rules I_R, I_R are A's institutional rules if and only if most people involved in A regard I_R as A's institutional rules.
>
> Given an institutional anchor A and a set of rules I_R, I_R are A's institutional rules if and only if (1) A's authority regards I_R as A's institutional rules and (2) most people involved in A defer to A's authority when it comes to determining which are A's institutional rules.

In a standard sense (see (Keefe, 2002), (Burgess, 2008), (Humberstone, 1997), (Boghossian & Velleman, 1989)) these two accounts are circular. Now, as I shall discuss at some length in Chapter 4, I do not believe that such circular accounts should be dismissed out of hand. For, their circularity notwithstanding, they can be insightful in many ways and, if the notion of institutional rules is a basic notion, non-liable to reductive analysis, they may be the best accounts we can come up with (again, for a more extant discussion, see Chapter 4).

Still, I take it that we should try to avoid circularity, for if there is a non-circular account of institutional rules available, it shall be superior to its circular alternatives. And the only way to find out whether there is such a non-circular account available seems to be trying to find one.

3.4. Accounting for institutional rules: power to enforce, commitment and publicity

Without further delay, I shall now present the account of institutional rules I'll be defending in this chapter.

Given an institutional anchor A and a set of rules I_R, I_R are A's institutional rules if and only if

(1) There is a group of people (the enforcers) who has the power to enforce I_R in A.

(2) A's enforcers are committed to I_R.

(3) A's enforcers' commitment to I_R is public.

(4) It is overtly believed in A that A's enforcers have the power to enforce the rules they are committed to in A.

I shall now discuss the different clauses of the account.

To start with: what is the power to enforce some rules, that clause (1) requires? As I shall understand it, the power to enforce some rules is basically a matter of power of discovery (the power to discover when rules are violated) and power of sanction (the power to sanction those who violate the rules). This suggests the following question: how strongly should we understand clause (1), i.e. how much

power of discovery and sanction should we take the enforcers to have? Obviously enough, we shouldn't understand it in a way that requires that the enforcers can discover and sanction every violation of A's rules. For, quite clearly, this isn't the case in most actual institutional contexts. Most contemporary states, for instance, don't have the power to catch all the people who commit fiscal fraud; and, similarly, most infractions to the traffic code are presumably never sanctioned because they aren't discovered by the relevant authorities. Hence, we should understand clause (1) in a more relaxed way (which I think is anyway more natural); roughly, in a sense in which a group of people has the power to enforce a set of rules if (a) in most of the cases, when they find out that someone has violated one of these rules, they have the power to sanction this person, (b) they find out at least sometimes when these rules have been violated and they could find out more often if they deemed it desirable.

Let us now turn to clause (2). This clause is built around the notion of commitment. As Searle stresses, "there are two components to the notion of commitment" (Searle, 2009, p. 81). Sometimes, we use it to convey the idea that someone (or some people) firmly intends to perform certain actions or achieve certain goals. In this sense, I might say "I'm committed to finishing my PhD thesis" to convey the idea that I really want to do so, I'm willing to do what it takes and so on. In some other context, we use the word commitments to refer to some kind of obligations. Hence, for instance, it makes sense to say "As a member of the police I'm committed to enforcing the law, but I can't fine everyone I see jaywalking… life would be impossible". And here, the idea that is being conveyed is something like: as a member of the police, it is my duty, function or role to enforce the law, but… My usage of the word commitment is closer to the second type of cases – for I want it to be possible for a group of people to be committed to certain rules even though they follow and enforce some of them only selectively, or even not at all. Hence, to be committed to Φ is, in the present sense, to be under some kind of obligation to Φ.

But not any kind of obligation is a commitment in my sense.[85] Hence, for instance, it may be a (moral) obligation of mine to help the poor – but, unless I signed up in an association whose aim is to help the poor I may yet fail to be committed to helping the poor.[86] And, if I intend to go to the supermarket afternoon, I may well be under some kind of rational pressure to take the necessary steps to this goal – but I'm not yet committed to do so, at least not until I've made this intention of mine public.[87] So what is commitment?

The notion of commitment I want is one according to which:

- The competent speakers involved in a linguistic community are (at least most of them) committed to follow the rules of the language they share
- Someone who signs a contract is committed to abide by the terms of this contract
- Someone who says "I promise to go to your birthday party" is committed to go the relevant Birthday party.
- When two people go for a walk together and there is no coaction involved, each of them is committed to do his part of their collective endeavor.

And the feature common to all these examples seems to be that, in each case, individuals acquire obligations in virtue of their participation into certain cooperative activities. Hence, I shall define commitment as follows:

> COMMITMENT: x is committed to Φ if and only if (1) x has the obligation to Φ and (2) x's obligation to Φ is grounded in x's participation in a certain cooperative endeavor.[88]

[85] I shall henceforth stop making explicit that "commitment" should be understood as "commitment in my sense".

[86] This doesn't imply that commitments aren't moral obligations, but only that not every moral obligation is a commitment.

[87] This doesn't imply that if I make my intention public I immediately acquires commitments, but only that I could acquire commitments by making my intention public in suitable circumstances.

[88] It is probably worth underlying at this stage that I do not think (or at least do not wish to commit myself to) the claim, defended in many places by Gilbert (e.g. (Gilbert, 1990) and (M. Gilbert, 2013)), that cooperation essentially involves obligations. Hence, COMMITMENT should not be understood as implying that there is a distinctive or irreducible kind of obligations essentially involved in cooperation, and that someone has a

Now, it isn't completely straightforward how we should understand the phrase "being committed to rules I_R". For, typically, commitments relate people to *actions* – hence we can be committed to *attend* the board's meetings, committed to *go running* every day, committed to *read* all the literature on collective action and so on… It may thus appear that the phrase "being committed to rules I_R" fails to specify a crucial bit of information (the nature of the actions one is committed to perform in relation to rules I_R) and that its content is therefore defective. But is it so? As matter of fact, we often speak of commitments to certain entities without specifying the actions required by such commitments. Hence, it makes perfect sense to say: "I'm committed to my relationship", "I'm committed to this team" or "by signing this contract, you commit yourself to the goals of our company". And, in these three cases, although no action is explicitly specified, it is clear that some *and only some* behaviors fit with the commitments in question. Hence, trying to save my relationship, contributing to the team's cohesion and helping to achieve the company's goals fit well with the three commitments mentioned above, while laughing at my relationship, disbanding this team and ridiculing the goals of our company do not. In other words, while phrases such as "being committed to this team" and "being committed to the goals of this company" may be elliptic – insofar as they fail to fill in the argument place corresponding to the actions which are the objects of the commitments in question – their content need not be defective, for they may still informatively constrain how this argument place is to be filled in. I believe that it is natural to understand the phrase "being committed to rules I_R" just in this way. Hence, a commitment to rules I_R may require that one follows rules I_R, or that one enforces these rules, or that one promotes them, etc… but it may not require that one systematically violates them, laugh at them or publicly despise them.

As for clause (3), it appeals to the notion of something being public. This notion has different usages too. Hence, when we speak of *public* schools we speak of schools funded by the government; when

commitment if and only if he has such an obligation. Rather, COMMITMENT presupposes that, sometimes, we acquire obligations by engaging in cooperative activities (which, for all I've said, may be *moral* obligations); and, with this presupposition in place, it defines "commitments" as the obligations we so acquire. Furthermore, it should be clear that COMMITMENT is intended as a stipulation, rather than an analysis of an ordinary notion.

we speak of *public* spaces, we mean spaces that are open to people; and when we say that the coach of the Barcelona Football Club made a *public* statement, we mean that he intended that his statement be generally accessible to people – i.e. that they could get to know that he made this statement. I use the word "public" in the sense of this later examples. Hence, clause (3) requires that the enforcers's commitment to rules I_R be *epistemically* accessible.

Finally, clause (4) appeals to the notion of overt belief discussed in Chapter 2 (p. 59.). Hence, it implies that (1) each of the people involved in A believe that the enforcers have the power to enforce the rules they are committed to in A; (2) that each of the people involved in A believe that (1); etc. Crucially, though, clause (4) doesn't require that the people involved in A know which are the rules the enforcers are committed to (that would violate the requirement that most people involved in an institutional anchor A may ignore which are A's rules), but only that they know that enforcers are committed to some rule.

With these clarifications in mind, I shall now bring support to the power to enforce-commitment-publicity account of institutional rules (henceforth PECP account of institutional rules). I'll have a three-way strategy to this end. Firstly, I'll motivate the main components of the account. Secondly, I'll show that the account satisfies the requirements and desiderata laid out in Section 3.3. Finally, I'll consider several – fairly different – institutional anchors and show how my proposal accounts for the fact that they have the rules they have.

3.4.1. Motivation for the different components of the account

As its name suggests, the PECP account of institutional rules has three main conceptual components. I shall motivate them one by one.

Let me start with the clause which requires that some of the people involved in the institutional anchor A have the power to enforce A's rules. Generally put, clause (1) is motivated by the intuitive thought that institutional rules must be supported by the possibility of sanctions. This is most obviously true of the law – since, quite clearly, there can be now law in the absence of a system whose function is to

sanction it violations. But this is also intuitively true of, for instance, the rules of universities (whose violations may be sanctioned in different ways, ranging from a simple reprimand by the universities' authorities, to temporary suspensions or even exclusion), the rules of religious communities (whose violation may be sanctioned by mandatory penitence activities or exclusion form the community), and even the rules of etiquette (whose violation may be sanctioned by social ostracism). This suggests that the existence of institutional rules must involve people who can discover such violations and sanction them, that is, people who have the power to enforce them.

To further motivate clause (1), let us consider the following story.

> Pau and I meet on Sunday. We are very angry about the impact Airbnb has on our neighborhood. We agree on a neighborhood chart which forbids Airbnb. We print gigantic posters which read "We, Pau and Aurélien, hereby forbid Airbnd in the neighborhood of Gràcia" and hang them all over the places.

Quite clearly, my friend Pau and I seem a bit nuts. Obviously enough, we cannot decide to forbid Airbnb in the neighborhood of Gràcia. And thus, obviously enough, the rule we agreed upon does not come to be an institutional rule of this area. But, let me ask, what went wrong? Why is it that Pau and I fail to create a rule for our neighborhood? According to the PECP account, our failure is due to our lack of power: we don't have the power to enforce the rule which forbids Airbnb in Gràcia and because of this we cannot turn it into an institutional rule. On the other hand, says the account, had we had the power to enforce this rule, then our public commitment to it would have made it an institutional rule of our district.

In order to turn this into an argument in favor of the power to enforce clause that is built into my account, I shall now argue, on the one hand, that the most salient alternative explanations of our failure are wanting and – on the other – that had we been much more powerful, we would indeed have been successful.

Let us first consider two salient alternative explanations[89] of our failure and show them wanting. According to a first explanation, the reason why we fail to create institutional rules for our neighborhood is that we don't have the standing to do so: we weren't elected, chosen or appointed to any position that would give us the right to make the rules of the Gràcia neighborhood, and thus we simply can't decide which rules shall be Gràcia's. Now the problem with this proposal is that it relies on the following principle: in order to choose the rules of an institutional anchor A, one must have the standing or right to do so. But, very plausibly, that one has such a standing is itself an institutional feature of reality. As such, it presupposes institutional rules and, therefore, any account of institutional rules containing this principle would fail by the light of the non-fundamentality requirement.

Some may be tempted to resist this objection by appealing to natural rights. Hence, they would argue, there are natural rights – which, by definition, do not presuppose institutional reality – and such rights lay at the foundation of institutional reality. I have one main quarry with this rejoinder: I do not believe that it tells a plausible story about the emergence of institutional reality. For it has it that at the (logical) beginning of institutional reality, there are people with the natural right to create certain rules. And hence, it implies that the birth of institutional reality must be somehow legitimized by an independent source of normativity. And I find this claim extremely untoward: as I see it, we have no reason to think that the cements of institutional reality must be somehow fair or correct by the light of any natural law. In other words, I believe that the attempt to ground institutional reality upon natural rights would fail by the light of the unfairness requirement of Section 3.3.[90]

According to a second explanation, the reason why Pau and I fail in our attempt to create new rules for the Gràcia neighborhood is that the people who live in the neighborhood do not come to regard

[89] They are salient, I take it, because they are naturally suggested by very influential theories on the market (see (Searle, 2009) and (Tuomela, 2003)).

[90] As far as I can see it, this isn't to say that social contract theorists (see (Rousseau, 2014), (Hobbes, 1982), (Locke, 1993), (Rawls, 1999)) are wrong: for their attempt to derive institutional rules from natural rights are better seen, I believe, as aiming at explaining how and which institutional rights could be legitimized by natural law.

our rule as a rule of the neighborhood. The problem with this proposal is that it is committed to the following principle:

- R is one of Gràcia's rules only if it is regarded as one of Gràcia's rules by the people who live in this neighborhood

And this principle would make the account fail, both by the light of the requirement that people can ignore the rules of the institutional anchor they are involved in, and by the light of the desideratum which commands to avoid circularity.

But what if Pau and I had had much more power: could we have been successful in our attempt to forbid Airbnb in the neighborhood of Gràcia. Suppose for instance that we are the bosses of a tremendously powerful gang based in Gràcia. Thus, when we publish our posters, it becomes clear to most inhabitants of Gràcia that anyone using Airbnb in the neighborhood will face serious consequences. They know how we proceed usually, so that they can imagine that the people we catch will first have to pay us great amounts of money, and then – in case of recidivism – may be harmed or even killed. Hence, they see our commitment to forbid Airbnb in Gràcia as a reason not to rent out their flat through this platform. In these circumstances, it is, I believe, correct to say that Airbnb is forbidden in the Gràcia neighborhood – or that we (Pau, myself and our gang) forbid Airbnb in this area.

To those who want to resist this later claim of mine, I'll ask to consider the case of the institutional rules of a tyrannical state – whose authority relies exclusively on the fear of its subjects. What could explain that, whereas Pau and I fail to create rules for our neighborhood, such a state succeeds in imposing rules on its territory?[91] I can think of only two explanations. On the one hand, some may

[91] Some may claim that, insofar as they are completely illegitimate, such states fail to create rules. They may be motivated by a Dworkinian understanding of value (see (Dworkin, 2013)) and law (which rejects legal positivism (see (Dworkin, 1977)). I do not wish to engage with such views, for as I see it, they do not have the same target as I do. That is: given the pre-theoretical conception of institutional rules I start with, it is a truism that illegitimate states can have rules. Hence, anyone denying this claim is better seen as aiming at accounting for a different phenomenon.

claim that – their illegitimacy notwithstanding, tyrannical states somehow have the right to choose the rules which apply in their territory, a right which Pau and I do not have; on the other, some may claim that – unlike Pau and I – tyrannical states get their subjects to regard the rules they adopt as holding (in the relevant territories). But, for reasons I've explained above, I think these explanations ought to be rejected, since they appear to violate the non-fundamentality, the non-circularity, the unfairness or the individual ignorance requirement.

One may further worry that Pau and I couldn't forbid Airbnb in Gràcia in this way since, obviously, pretty much all the actions of our gang would be illegal. But this, I take it, would just be a case of institutional conflict. Institutional conflicts are quite a mundane phenomenon: for instance, many states have laws which overtly conflict with the laws of the international organisms they belong too. And by no means would it seem reasonable to infer that these (internationally illegal) laws aren't institutional rules of the states in question. Likewise, I don't think that the illegality of the actions of the gang Pau and I lead threaten the fact that these actions impose new rules in the Gràcia neighborhood.

Let me now turn to the commitment condition, expressed in clause (2) of the account. It is, I believe, easy to see why an account of institutional rules based exclusively on power would be wanting. For the fact is that power needs *not* be tight to a particular set of rules (although it may be). Consider a small tribe whose rules are I_R. Suppose furthermore that the chieftain's family is extremely powerful in this tribe – so powerful indeed that they could enforce not only I_R, but many different sets of rules. What makes it the case then that the tribe's rules are I_R rather than some other rules? The commitment condition offers an answer: the chieftain's family is committed to I_R, rather than any other set of rules.

But is the claim that the enforcers are committed to A's rules the right way to complement a power based account of institutional rules? I think it is.

To start with, I think it is intuitively very plausible that, whenever an institutional rule is in place, at least someone must have some kind of obligation to somehow contribute to making it the case that

the rule is followed (by either following the rule, or promoting it, or enforcing it, etc.). And, given our discussion of the notion of commitment to a rule, condition (2) ensures that this intuition is accounted for by my account.

Now, obviously, the intuition I appeal to above is so undemanding that it may be accounted for in countless different ways. Hence, the support (2) inherits from the above reasoning is meager. More explicitly, (2) requires:

- That the contributive obligations the intuition refers to be had by the enforcers
- That these contributive obligations be commitments, i.e. grounded in the enforcers participation in cooperative activities

And the intuition we started with does bring any support to either of those claims. In order to strengthen my case, thus, I shall now discuss two examples in detail:

> *States*: in any state S, the enforcers are, presumably, either civil servants or elected politicians. And in both cases, they seem to acquire contributive obligations to S's rules in virtue of the fact that they voluntarily come to occupy an official position which is essentially tight to these rules. But that they voluntarily come to occupy such a position is the result of a cooperative activity. For it takes at least to people to make it the case that someone comes to voluntarily occupy an official position: one to want this position and one to attribute it. And if the position is finally filled in, they must cooperate in making this happen. Hence, in any state S, the enforcers are plausibly seen as committed to S's rules.
>
> *Money rules*: consider the following story. The Hoppers use a certain type of sea-shells as money (call it SM). The rules which underlie this behavior of theirs were never explicitly agreed upon, they just developed out of the disposition the Hoppers had to barter goods and services for SMs, which they find very beautiful. In such a scenario, the Hoppers (all together) have the power to enforce those rules – hence they could, in most of the cases,

punish one them who would have fraudulently use SRs (a different – but apparently very similar – kind of sea-shells) as money; furthermore, the Hoppers seem to be each committed to the sea-shells money rules by their daily monetary transactions. But such transactions are clearly instances of collective activities. Hence, in this scenario too, the enforcers (i.e. in this case the whole group of people involved in the institutional anchor) appear to be committed to the anchor's rule.

Now, I take it to be relatively clear that the kind of reasoning I applied to these cases shall smoothly generalizes (hence, in a religious institution (e.g. the catholic church), the enforcers are, plausibly, people who occupy an office in the institution (e.g. priests, bishops, archbishops, etc.) and these people have contributive obligations with respect to the rules of the church because they voluntarily came to occupy the relevant positions; in a schoolyard where there is a rule to the effect that every kid should show proper respect to Alberta, the boss of the yard, the enforcers are plausibly seen as the group of bullies surrounding Alberta – and they appear have an obligation to enforce the above rule because of their voluntarily affiliation to Alberta's group; etc.). Thus, the claim that the enforcers of an anchor A are committed to A's rules (i.e. clause (2)) appears to be in good standing.

Furthermore, I take the following observations to give additional reasons to consider that the notion of commitment to a rule aptly captures the relation between the enforcers and institutional rules:

- As per the unfairness requirement, this relation shouldn't presuppose that institutional rules are good rules – and clearly one can be committed to rules that aren't good.
- As per the non-circularity requirement, it shouldn't be the case that standing in this relation to a rule presupposes that one sees this rule as an institutional rule – and clearly one can commit to a rule without seeing it as an institutional rule.
- As per the non-fundamentality requirement, it shouldn't be the case that standing in this relation to a rule presupposes the existence of institutional reality – and although commitment to rules is sometimes grounded in pre-existing institutions (e.g. I sign a

contract with a university and hereby becomes committed to its regulation), it is presumably not necessarily so. Hence, it seems that, even in the absence of pre-existing institutions, a group of people could commit to a set of rules by agreement (this is, to be fair, a complex issue – which I'll address in more detail later on, when arguing that my account meets the requirements of Section 3.3).

- Enforcers may relate to the rules of the relevant institutional anchors in different ways. Sometimes, they are part of the collective the rules are meant to apply to (e.g. civil servants with respect to the law, university staff with respect to the university's regulation, etc.), in which case they are likely to be committed (among other things) to follow the rules. But sometimes they aren't subject to the relevant rules (e.g. the board of some universities have members who aren't staff, in some kingdoms the king is above the law, etc.), and, in these cases, they aren't committed to follow the rules in question, but rather to enforce or promote them. And the notion of commitment to some rules has the flexibility necessary to accommodate these different situations.

I shall now turn to clause (3), which requires that satisfaction of conditions (2) be public. The reasoning backing up this condition is the following: the function of the rules of an institutional anchor is to drive the behavior of the people involved in it. And this function cannot be performed unless the people whose behavior they are meant to drive know them. Hence, it appears that the ideal functioning of institutional rules would require universal knowledge (and presumably even common knowledge). But, because of the obvious fact that most people involved in an institutional anchor can ignore many of its rules, we cannot require universal knowledge. What we can require though, compatibly with the possibility of massive individual ignorance, is that such knowledge be accessible, i.e. that institutional rules be public. But now, as discussed above (while motivating clause (2)), it is the commitments of the enforcers that determine which are the rules of a given institutional anchor. Hence, if I_R are A's institutional rules, the piece of information which must be public is that A's enforcers are committed to I_R. Thus clause (3).

Let me finally turn to clause (4). The motivation for the clause is the intuitive thought that institutional reality cannot exist unless the people it affects are minimally aware of it. In more explicit terms, I find it intuitively compelling that the people involved in an institutional anchor equipped with institutional rules must believe this much: that they are involved in social contexts in which some rules apply. Now, clause (4) is my proposal to reconcile this intuitive thought with the uncontroversial observation that people may ignore which rules they are supposed to follow in the different institutional anchors they are involved in. For, given the recursive nature of overt belief (see Chapter 2, p.65), clause (4) ensures that the people involved in an institutional anchor which has rules believes it, while still leaving open the possibility that they may ignore which are the rules of this institutional anchor.[92]

3.4.2. The account satisfies the requirements and desiderata of Section 3.3.

The PECP account of institutional rules clearly satisfies most of the requirements and desiderata of Section 3.3.

Individual ignorance of institutional rules: the account makes room for the possibility that most of the people involved in an institutional anchor A fail to know which are its rules. Firstly, because it only requires that some of the people involved in A be committed to A's rules – while the others must only be in a position to acquire knowledge regarding A's rules. Secondly because even (some of? maybe all?) the people committed to A's rules, may ignore which are the rules they are committed to. To see this, consider the case of someone who is hired by a company and, in the process, signs a contract which stipulates that she accepts the company's regulation. By doing this, she commits herself to the company's regulation but she may well ignore its content – if, for instance, she hasn't read it.

[92] Indeed, we should be careful not to interpret clause (4) in a way that implies that every person involved in the institutional anchor should know who the enforcers are (a part, obviously, under the trivial description: the people who have the power to enforce A's rules). For I believe this would be too much individual knowledge already. On the other hand, I do think that we should understand (4) in a way that entails that the people involved in an institutional anchor A believe that they are involved in A. For I think that the existence of institutional rules requires that there be groups of people which meet the following self-awareness condition: group G is self-aware if and only if (1) everyone in G believes that she belongs to G, (2) (1) is overtly believed in G. Notice that this is a very weak notion of self-awareness, according to which, for instance, the group of blond people is likely to be self-aware.

Individual violation: the account obviously makes room for the possibility of massive violation of institutional rules.

Individual dissatisfaction: the account allows that most of (or maybe even all) the people involved in an institutional anchor A be dissatisfied with its rules. This is obvious in the case of the people who aren't committed to these rules. But this is true, even though maybe less obviously, of those people who are committed to them: for it seems that I could commit myself to the rules of a company – motivated, say, by the salary I shall receive – even though I find these rules stupid or unfair.

Unfairness: nothing in the PECP account of institutional rules even remotely suggest that such rules should be fair, just or morally good in any sense.

Selective enforcement or absence thereof: it is less than obvious that the PECP account requires that the rules of an institutional anchor ever be actually enforced. At any rate, it clearly makes room for selective enforcement.

Avoiding circularity: it seems quite plausible that the PECP account of institutional rules avoids circularity. For the main components of the account are, at least prima facie, conceptually independent from the notion of institutional rule. Hence the notion of enforcement appears to involve the idea of checking whether a rule is followed or not, together with the idea of punishing or sanctioning violations thereof, and none of these notions seem to have the concept of institutional rules built into them. As for the notion of power, it is certainly extremely complex. But the following seems uncontroversial: in some circumstances, power can be entirely grounded in physical strength (or, for what matters, intelligence) – and this appears to exclude that the notion of power be conceptually dependent upon the notion of institutional rules. The notion of publicity shouldn't generate much trouble either, since it is to be defined in terms of epistemic accessibility, and the same goes for the notion of overt belief. This leaves us with the notion of commitment. This last notion is, I believe, the only one that may be suspected of conceptual dependence on the notion of institutional rules. I can't fully set this worry aside here: for it would require that I put forward a complete analysis of this notion – and this clearly

goes beyond the scope of this chapter. Thus, I shall stick to this claim, which seems uncontroversial enough: the account I defend doesn't look circular and if it is, the analytic circle it contains is big, big enough indeed to ensure that the account is philosophically illuminating (for more on illuminating circular accounts, see Chapter 4).

Non-fundamentality: last but not least, I believe that the PECP account of institutional rules meets (well enough) the non-fundamentality requirement.

As discussed in Section 3.3, this requirement demands that an account of institutional rules suggests a story about how institutional reality emerges from other, intuitively more fundamental, aspects of reality. It is pretty clear which is the story that the account I defend in this chapter suggests. According to this story, there are institutional rules when:

(1) some people, by engaging in cooperative activities, come to be committed to certain rules which they have the power to enforce.
(2) the commitments in (1) are public
(3) there is a suitable degree of mutual awareness regarding (1)[93]

Hence, whether the account meets the non-fundamentality requirement or not depends on whether power, commitments to rules, and publicity are aptly described as more fundamental than institutional reality.

I believe that publicity shouldn't cause any trouble in this respect. For, as I use the notion, publicity is a matter of epistemic access – i.e. that something is public in a population P roughly means that P's members can get to know it. And epistemic accessibility is, at the very least, not less fundamental than institutional rules (i.e. it isn't the case that whenever something is epistemically accessible, this is so

[93] What's a suitable degree of mutual awareness regarding (1) is discussed in more details above. (See footnote 92).

partly in virtue of the fact that some institutional rules are in place). And, by the same reasoning, I take it that the mutual awareness requirement is no threat.

The appeal to power shouldn't generate much trouble either. For, even though some power is indeed institutionally grounded – in the sense that the people who have it, have it (partly) because of the institutional rules that are in place in the social contexts the belong to –, there are other, intuitively more fundamental, sources of power: hence, in some contexts, sheer physical strength can make someone powerful.

The matter is more complicated when it comes to commitments to rules. For the claim that commitments to rules presuppose institutional reality – and is therefore by no means more fundamental – doesn't seem too far-fetched. I've defined commitments as follow:

> COMMITMENT: x is committed to Φ if and only if (1) x has the obligation to Φ and (2) x's obligation to Φ is grounded in x's participation in a certain cooperative endeavor.

I take it to be clear that cooperation doesn't presuppose institutional reality. At any rate, this seems to be the case if the account I've offered in Chapter 2 is correct. But still, one may worry that the obligations in (1) presuppose institutional reality; and there is also room to argue that, even though commitments don't require institutions, commitments *to rules* do.

To frame our discussion, consider the following story:

> A bunch of colonists separately reach the Yodo territory. The Yodo territory really is a no man's land. On the day of their arrival, the colonists meet and agree on a basic chart that will govern their life in Yodo.

Now, a superficial appraisal of this scenario may suggest that the colonists have adopted a set of institutional rules – (partly) by committing themselves to these rules – with no institutional background contributing to ground their commitments. But the matter is clearly not so simple. For it is nothing but natural to consider that, in reaching an agreement regarding the rules that shall govern their life

together, the colonists have been using language. And the literature has traditionally regarded language as an institutional feature of reality[94]. Hence Searle says:

> "One reason for the inadequacy of the tradition [that studies institutional reality] is that the authors, stretching all the way back to Aristotle, tend to take language for granted. They assume language and then ask how human institutions are possible and what their nature and function is. But of course, if you presuppose language, you have already presupposed institutions. It is, for example a stunning fact about the Social Contract theorists that they take for granted that people speak a language and then ask how these people might form a social contract. But it is implicit in the theory of speech act that, if you have a community of people talking to each other, performing speech acts, you already have a social contract. The classical theorists, in short, have the direction of analysis back to front." (Searle, 2005, pp. 1–2)

Hence my account faces the following charge: commitments to rules presuppose language; and language is itself an institutional feature of reality; therefore, the PECP account of institutional rules presupposes institutional reality and is therefore unfit to explain how it emerges from more fundamental aspects of reality.

I have two replies to this worry. One is cautious and relatively uncontroversial; the other is more ambitious and, accordingly, more controversial.

The cautious reply grants that commitments to rules presuppose language and that language belongs to institutional reality, but, it goes on to insist, the PECP account can still be seen as a satisfactory account of institutional reality minus language (henceforth, non-linguistic institutional reality). That is, according to the cautious reply, the PECP account should be seen as aiming at accounting for institutional rules concerning judges, money, borders, doctors, proper manners, etc. but not for the rules of language, which determine that "frog", "red", etc. mean what they do. In other words, the PECP account aims at telling a story about how people, once they have language, can impose non-linguistic institutional statuses onto the world around them, while leaving the story about the emergence of linguistic meaning to a different chapter of philosophy.[95]

Now, quite clearly, whether the cautious reply is successful depends on whether language presupposes what I've labelled non-linguistic institutional reality. If it doesn't, then it can be rightly seen as more

[94] See (Searle, 2009) and (Tuomela, 2003), for two clear examples.
[95] Given that I've avoided to use linguistic examples to introduce the target of my investigation, I would even argue that it isn't obvious that the cautious reply changes topic.

fundamental than the latter, and the PECP account of non-linguistic institutional reality meets the non-fundamentality requirement. On the other hand, if language does presuppose non-linguistic institutional reality, we are back on the loop we were trying to avoid – and my account still fail by the light of the non-fundamentality desideratum.

How plausible, then, is the claim that language doesn't presuppose non-linguistic institutional reality?[96] I don't know – and I would rather avoid having to answer such a deep and elusive question to salvage my account. One thing is sure though: there are many paradigmatic examples of institutional phenomena that language doesn't presuppose – money, borders, state, governments, judges, companies, universities, etc. And these phenomena constitute by themselves a target worth investigating. In particular, an account making explicit their commonalities and distinctive traits, and explaining how they are grounded in more fundamental aspects of reality would certainly be worth having. Therefore, there seems to be a cautious way to conclude my cautious reply: instead of the elusive "institutional reality minus language", define non-linguistic institutional reality as that bit of institutional reality which isn't presupposed by language.

To sum up then: the cautious reply to the non-fundamentality worry, (1) grants that language belongs to institutional reality, (2) grants that commitments to rules presuppose language, (3) proposes that we see the PECP account as an account of non-linguistic institutional reality (understood as that bit of institutional reality which isn't presupposed by language), and (4) insists that such an account is worth having, for it reveals the commonalities between many phenomena (money, judges, borders, state, etc.) that have been of central concern to most people studying institutional reality, as well as it

[96] For what is worth, Searle thinks it is a platitude:

> "It is intuitively obvious, even pre-theoretically, that language is fundamental in a very precise sense: you can have language without money, property, government, or marriage, but you cannot have money, property, government, or marriage without language." (Searle, 2005, pp. 11–12)

And this at least puts the cautious reply on dialectically safe ground vis-à-vis those who object to my account on the basis of Searlian considerations on the role of language in the creation of institutional reality (see (Searle, 1995 Ch.3)).

explains how such phenomena emerge from more fundamental aspects of reality (in particular, linguistic institutions).

The more ambitious reply denies that commitments to a rule requires language. Now, given the extreme complexity of the issue at stake, and given the availability of a less committal – although less satisfactory - reply to the worry I'm facing, I shan't try to offer an elaborated, definitive answer. This notwithstanding, I submit the following story, which – I take it – makes a good prima facie case:

> *Herbert and Alfred*: Herbert and Alfred belong to two tribes which inhabit adjacent territories. The two tribes have different languages[97] and neither Herbert nor Alfred speak each other's language. Both Herbert and Alfred enjoy spending time on the shore of a river which runs between the tribes' territories. One day, they come across each other by the riverside. At first, they are rather distrustful, but they soon relax and, at some point, they start throwing stones into the river together and enjoy it a lot. Then they leave. The story repeats itself more and more often and, at some point, Herbert and Alfred meet every day around 6PM on the shore of the river to throw stones into the river together.

I find it plausible that, after their meetings by the riverside have been a regular thing for some time, Herbert and Alfred come to be committed to showing up at the right time and place, and throwing some stones into the river together. More precisely: I think that, in this story, (1) a Lewisian convention has emerged (see (Lewis, 1969)), (2) Herbert and Alfred came to be committed to follow the corresponding rule. And I fail to see how this presupposes language in any sense.[98]

[97] In a broad sense, which includes gestures endowed with a conventional meaning.

[98] A possible reply is that, prior to the emergence of the Lewisian convention, Herbert and Alfred must have developed a language of their own (or that one must have taught pieces of his language to the other). I don't think this is very plausible, though. Another possible reply is that, if they hadn't grown in a linguistic environment, Herbert and Alfred couldn't have had the conceptual apparatus needed for the story to unfold as told. But, once again, I take it that the claim my objector is submitting is, on the face of it, rather controversial.

3.4.3. Some illustrations

To conclude my defense of the PECP account, I shall now consider different institutional anchors and show how my proposal account for the fact that they have (some of) the rules they have.

> *Poetry reading group:* a group of friends decide that they would want to read and comment poetry together. They meet and, together, choose the rules which will govern the existence of the group: meetings will be held on Friday, in a Bar next to the train station and the sessions will be guided by a president, elected every three months by a majority vote of the members of the group. New members shall be accepted by a majority vote of the members of the group.

According to the account of institutional rules I'm defending in this chapter, the poetry reading group has the rules it has because: its members are committed to its rules (the founding members are committed by the explicit agreement they reached; the other members are implicitly committed by the fact that they asked to enter a group already equipped with a set of rules); its members have the power to enforce the group's rules (by, for instance, excluding from the group those who fail to abide by them); and all of this is overtly believed by the group members (and hence public).

> *Terrorist tax*: an armed group – deemed illegal by the state which rules over the territory it inhabits – forces all the inhabitants of area B to pay a terrorist tax of 100 euros per month. The rule is clear and has been publicly expressed by the group in posters and declarations in the media.

In this case, my account says that it is a rule of area B that its inhabitants have to pay the terrorist tax for: the members of the armed group (or at least those who detain the power in it) are committed to the tax rule because they've jointly chosen it; the armed group has the power to enforce this rule; this situation is common knowledge among the B inhabitants.

> *The University's Regulation*: R is University U's regulation.

According to the PECP account, this is so because: most of the people involved in U are committed to R (U's employees because they (implicitly or explicitly) agreed to U's regulation when they signed their contract, U's clients (mainly the students) because they (again implicitly or explicitly) agreed to U's regulation when they 'hired U's services'); some of the people committed to R (the University's board, or the dean, etc.) have the power to enforce the University's regulation (by sanctioning, firing or excluding those who violate it); the commitment of U's enforcers to R is public in U – its members can get to know which are rules U's enforcers are committed to by reading U's regulation; and the people involved in U overtly believe that U involves some people with the power to enforce in U the rules they are committed to.[99]

Saying "Hi" in a gesture: The Hoofers say "Hi" by executing a particular gesture (call it A).

The rules which underlie this behavior of theirs were never explicitly agreed upon.

According to the PEPC account, the rules which underlie the A-greetings are rules of the Hoofers because: most of the Hoofers are committed to these rules by the fact that they rely on them in their daily interactions; the Hoofers have the power to enforce those rules – hence they could, in most cases, ostracize someone who would fail to use the A gesture to say "Hi", or would systematically use a different gesture for this purpose; all of this is public to the Hoofers; all of this is overtly believed by the Hoofers.

[99] Here, some may worry that in explaining how the University's employees and clients get to be committed to U's regulation I have presupposed that R is U's regulation. They are right about the fact that I've made this presupposition but wrong – I believe – to be worried about it. True: it is because R is U's regulation at time t that, when Joaquin signs a contract with U, he becomes committed to R. True also: if Joaquin is among U's board, it is partly because Joaquin is committed to R that, at any time t* between t and the moment in which Joaquin stops working for U, R is U's regulation. But there is no explanatory loop here, once we make clear the different time components of the relevant facts. Fair enough: the story about how the first members of the University's board came to be committed to R might have to be a different one (for, presumably, R wasn't U's regulation before they became U's first board's members): but this is no reason to worry either – the first board members may have acquired their commitment by explicit agreement (like the founding members of the poetry group reading above), or because they accepted to become the first board members of a University-to-be for which a regulation was already written (maybe by the state, in the case of public a university, maybe by a private sponsor).

Chapter 4: Circular accounts of collective phenomena: a discussion

Abstract: in this chapter, I argue for the modest and yet important claim that – their limitations notwithstanding – circular accounts of collective phenomena can yield interesting insights and answer pressing questions about these phenomena. I vindicate the relevance of this claim by arguing that – once it is taken into account – Searle's argument for his extravagant conception of collective intentionality doesn't even get off the ground. I conclude by suggesting that – under not too farfetched assumptions – circular accounts of collective phenomena may be the best we can come up with.

4.1. Introduction – circular accounts of collective phenomena and their detractors

Most philosophers who study collective phenomena have been trying to put forward conditions which are *a priori* sufficient (see, for instance, (Bratman, 1993)) or both necessary and sufficient (e.g. (Kutz, 2000), (Gilbert, 1990), (Tuomela & Miller, 1988)) for their occurrence. Let us refer to such set of conditions as *accounts of collective phenomena*. Some of these proposals are circular: in one sense or another, the notion they target (or a close inter-definable relative thereof) is used in stating the conditions which do the accounting. A paradigmatic example of such circular accounts of collective phenomena (henceforth CACPs) is Margaret Gilbert's account of plural subjects:

> "Consider again the complex logically necessary and sufficient condition for the existence of a plural subject. This is (to repeat) that a set of persons with the concept of a plural subject must have openly mutually expressed their willingness to be members of such a subject, and this is common knowledge." (Gilbert, 1992, p. 205)

Many social ontologists have expressed skepticism about the value of CACPs. For instance, in discussing a purported counterexample to Tuomela and Miller's account of collective intention, Searle says:

> "Could we avoid such counterexamples by construing the notion of "doing his part" in such a way as to block them? I think not. We are tempted to construe "doing his part" to mean doing his part toward achieving the *collective* goal. But if we adopt that move, then we have included the notion of a collective intention in the notion of "doing his part." We are thus faced with a dilemma: if we include the notion of collective intention in the notion of "doing his part," the analysis fails because of circularity; we would now be defining we-intentions in terms of we-intentions. If we don't so construe "doing his part," then the analysis fails because of inadequacy."(Searle, 1990, p. 405)

On a different line, Björn Petersson has argued that circular accounts are unfit to fulfill the main purpose of the analysis of collective phenomena:

> "I suppose that the main theoretical purpose of the analyses at hand [i.e. the analysis of collective action and intention] is to establish a substantial distinction between collective actions and noncollective sets of individual actions. Circularity clearly threatens that aim. So, in this case, analytical circularity is a methodological problem, and a substantive one."(Petersson, 2007, p. 146)

And Michael Bratman explicitly tries to avoid circularity in his constructive account of collective intentions:

> "my strategy is to see our shared intention to *J* as consisting primarily of attitudes of each of us and their interrelations. At least some of these attitudes will specifically concern our joint action of J-ing; after all, our shared intention to *J* supports coordination specifically in the pursuit of our *J*-ing. But much talk of joint action already builds in the very idea of shared intention. For us to try to solve a problem together, for example, we need an appropriate shared intention. We would risk criticisable circularity if our analysis of shared intention itself appealed to joint-act-types that involved the very idea of shared intention. So we will want to limit our analysans to joint-act-types that are, as I will say, neutral with respect to shared intention." (M. E. Bratman, 1993, p. 101)[100]

Overall then, circularity has been of great concern for many philosophers working on collective phenomena.[101]

In this chapter, I argue for the modest claim that, their limitations notwithstanding, CACPs can be very useful when one sets out to investigate the nature of collective phenomena. In particular, I show that CACPs may contain precious insights, and answer pressing questions about these phenomena. Hence, regardless of whether CACPs can be ultimately improved on by accounts which preserve their virtues while avoiding their flaws, we should be careful not to dismiss them too swiftly.

My plan is as follows. In Section 4.2, I highlight what I take to be some of the major limitations of CACPs. In Section 4.3, I point at their potential virtues. In Section 4.4, I suggest that Searle might have been driven to accept his extravagant conception of collective intentionality by ignoring the insights that CACPs provide with – hence vindicating the relevance of such insights. In Section 4.5, I offer some concluding remarks and briefly suggest that we shouldn't assume that CACPs can be improved on.

[100] The need to avoid circularity is also mentioned in Bratman's paper on shared cooperative activities ((Bratman, 1992, p. 330).

[101] Further evidence for this claim is provided by that fact that, if you type "circularity" in the Stanford Encyclopedia of Philosophy browser, the fourth result you get is an entry on collective intentionality.

Before starting with the core of my discussion, though, I must clarify a (partly) terminological issue. I've said that Gilbert's account of plural subject is a paradigm of CACP. Now, in her account, the occurrence of the target notion which is supposed to make the account circular is embedded in an intensional context. And, indeed, virtually all the accounts of collective phenomena whose 'circularity' is under discussion are deemed circular in virtue of an occurrence of the target notion (or a close relative thereof) in the scope of a verb of attitude (such as "desire", "intend", etc.).

Hence, in our discussion, we will want not only accounts relevantly similar to (1) but also those which suitably resemble (2) and (3) to qualify as circular:

(1) x is a city if and only if x consists of many agents who jointly constitute a city

(2) x is money in population P if and only if the members of P intend to use x as money

(3) the xs cooperate in J-ing if and only if each x intends to J by cooperating with the xs

The difficulty is that it isn't obvious how we should define such a notion of circularity.[102] We may want to try out a liberal proposal, according to which an account is circular if and only if its right-hand-side (somehow) *contains* the term that is being accounted for. But some shall worry that this conception overgeneralizes: the right-hand-side of a biconditional can *mention* the term under analysis, they'll argue, without hereby being guilty of circularity. Hence, we may be tempted by a more restrictive understanding, according to which an account is circular if and only if its right-hand-side *uses* (as opposed to merely mentions) the term that is being accounted for. But some will then complain that – on this conception – (2) isn't circular: they'll claim that the "that"-clauses introduced by verbs of attitude (such as "believe", "think", "intend", etc.) refer to the sentences they embed (e.g. in "the members of P believe that x is cool", "that x is cool" refers to "x is cool") which are therefore mentioned rather than properly used.[103]

[102] It is worth saying though, that reckoning not only (1) but also (2) and (3) as circular appears to be pretty standard. See (Keefe, 2002), (Burgess, 2008), (Humberstone, 1997), (Boghossian & Velleman, 1989, p. 89).

[103] This *self-referential* conception of "that"-clauses is sometimes labelled *sententialism*, after Schiffer (see (Schiffer, 1987)). Notice, nevertheless, that not every version of sententialism has it that the terms within the scope of verbs of attitude are merely mentioned. Higginbotham, for instance, claims that "the words of the

Fortunately, I need not try settling this issue here. For it is a fact that many authors who work on collective phenomena take issue with accounts such as (2) and (3) on the ground that the term under analysis occurs in the right-hand-side of these biconditionals (see (Bratman, 1993, p. 101), (Searle, 1990, p. 405) and (Petersson, 2007, p. 146)). What I want to know is the extent to which these worries are well motivated and not whether these authors are right to describe them as 'circularity worries'. I shall thus simply accept that – on an appropriate conception of circularity, whatever it is exactly – accounts such as (2) and (3), and not only (1), are CACPs. Then I'll argue that, at least for some purposes, CACPs so understood can be useful indeed.

4.2. Where CACPs fail

There are certain goals that CACPs cannot achieve, and certain purposes that they are unfit to serve. And even though I'm arguing in this chapter that CACPs shouldn't be generally and swiftly dismissed, it is important to also briefly go through (some of) their limitations, in order to get a balanced overview.

4.2.1. Conceptual reduction

The main limitation of CACPs is obviously that they are not reductive analyses; i.e. that they do not define collective concepts in non-collective terms, and fail to reduce collective notions to other concepts. For in any CACPs, the target notion – or another collective notion, inter-definable with the target notion – occurs in the right-hand-side of the biconditional, i.e. the would-be *definiens*.

4.2.2. Settling issues of conceptual primitivity

As they fail as reductive analysis, CACPs are unhelpful when it comes to deciding whether collective notions are basic or not. On the one hand, the truth of a CACP obviously doesn't show that the notion it targets is reducible. On the other, it doesn't show either that the target notion is basic: presumably, a CACP could be true, even though there is a reductive analysis available. For instance, the trivial biconditional "cooperation occurs if and only if cooperation occurs" is certainly true – and yet only a

complement clauses to which [sententialism] applies (…) are both mentioned and used" (Higginbotham, 2006, p. 110).

fool would think of trying to conclude from this that the notion of cooperation is basic. Hence, insofar as one is interested in knowing whether collective notions are primitive or not, one shouldn't look for an answer in CACPs.

4.2.3. Fully explicating the content of collective notions

CACPs also fail to fully unfold, reveal or make explicit the content of collective notions. To see this, consider the following circular and schematic account of cooperation:

(C) the Xs cooperate if and only if (a) each X intends to cooperate with the Xs, (b) BLA[104]

Assuming that (C) is a conceptual truth, it does tell us some things about the content of the notion of cooperation.[105] But it doesn't make explicit all of it. For, given condition (a), it is part of the notion of cooperation that those who cooperate hold certain intentions. But, in (a), the content of these intentions is specified thanks to the very notion of cooperation. Hence, (C) only specifies the content of the intentions distinctively involved in cooperation by drawing on… the notion of cooperation. Therefore, some aspect of this notion – which distinguishes an intention to cooperate from other closely related intentions – isn't unfolded or explicated in (C).

Another way to put the same point: (C) cannot make explicit everything one must know in order to understand the notion of cooperation. For, according to (C), we understand this notion only if we understand what an intention to cooperate is. But if we try to use (C) to explicate what an intention to cooperate is, we end up explaining intentions to cooperate in terms of… intentions to cooperate. Hence there is an aspect of our understanding of the notion of cooperation – that which allows us to grasp the distinctiveness of an intention to cooperate – which (C) presupposes rather than explicate.

[104] Where "BLA" stands for an indeterminate number of interesting, non-circular conditions.
[105] I'll go back to this in Section 4.3.1.

4.3. Where CACPs succeed

Despite their failures in other respects, CACPs can be helpful for many purposes. In particular, I'll now argue that they can give important insights into the nature of the phenomena they target (3.1) and that they can even individuate them in a non-trivial sense (3.2).

4.3.1. Insights into the nature of collective phenomena

It seems pretty clear to me that circularity by itself doesn't prevent an account of a collective phenomenon from giving us insights into the nature of the phenomenon it targets. Assume, for instance, that the following biconditional is true in virtue of the nature of cooperation:

(C1) Agents Xs cooperate in G-ing if and only if

(a) each X intends to cooperate with the Xs in order to achieve G

(b) each X acts in accordance with (a)

(c) (i) each X believes that (a) and (b) obtain, (ii) each X believes that (i), etc.

This biconditional contains many interesting pieces of information.

- *Common goal*: C1 implies that cooperation requires a common goal in the following sense: whenever some agents cooperate, there is a goal that each of them wants to achieve.
- *Mutual beliefs (or common knowledge)*: C1 says that cooperation requires some kind of hierarchy of mutual beliefs.
- *Cooperative behavior vs parallel individual actions*: Overall, C1 (at least) strongly suggests that cooperative behavior is made of individual actions performed with suitable intentions. Hence, what distinguishes cooperative behavior from parallel individual actions (or mere sums of individual actions) are the intentions which drive the agents who cooperate. And, more explicitly, the intentions of those who cooperate are characteristic because they involve the notion of cooperation.

- *Supervenience*: C1 seems to imply that cooperation supervenes on facts about individual agents – their mental states, actions and some quite specific relations they may stand in to one another. For according to C1, whether cooperation occurs, between whom, and to which goal, all of this can change only if some agents have different properties or stand in different relations to one another.
- *Collective intentions*: define collective intentions as the kind of intentions involved in cooperation. C1 strongly supports the claim that collective intentions are individual intentions, distinctive in virtue of their content (they contain the notion of cooperation) as opposed to, for instance, their subject (i.e the entity who holds the intention), their mode/form (as defended by Searle (see (Searle, 1990)) or the kind of reasoning which led to their formation (as defended by (Gold & Sugden, 2007)).
- *Reduction*: C1 strongly supports the claim that cooperation reduces to sums of individual actions. On the one hand, it suggests that the property "being cooperation" is identical to the property "being a sum of individual actions which satisfies conditions (a), (b) and (c)" (type-identity). On the other, it backs up the claim that every instance of cooperative behavior is a sum of individual actions suitably lumped together by the intentions with which they are performed (token-identity).
- *Self-referentiality*: C1 entails that cooperation is self-referential, i.e. each instance of cooperation involves agents who possess the notion of cooperation and represent what they are doing as cooperative behavior.

The list could be longer, but I think the point has been made: circular accounts of collective phenomena need not be trivial. They may have many interesting and controversial implications about the nature of the phenomena they target. Hence, (true) CACPs can provide great insights into the nature of collective phenomena.

There is one lingering worry though. Even if we grant that true CACPs can be insightful, we may still hold that CACPs are never insightful because they couldn't possibly be true. According to this line of thought, the kind of circularity involved in CACPs isn't one which triggers triviality, but rather inconsistence. However, there is no obvious incoherence in claiming that those who cooperate must intend (or desire) to cooperate – and hence the burden is on the defenders of the inconsistence thesis to provide with an argument that would back up their position.

They may think that an argument laid out by Boghossian and Velleman in their *Colour as a Secondary Quality* (Boghossian & Velleman, 1989) fits the bill. In this paper, Boghossian and Velleman criticize the dispositionalist theories of color properties which have it that, for instance:

(iii) Red $=_{df}$ a disposition to look red under standard circumstances

They claim that, on the assumption that "red" has the same meaning on both side of the definition symbol, (iii) is incoherent insofar as it "precludes visual experience form telling us which color an object has"(Boghossian & Velleman, 1989, p. 89). Their argument runs as follow:

> "Under the terms of (iii), an experience can represent its object as red only by representing it as disposed to produce visual experiences that represent it as red. The problem here is that the experiences that the object is thus represented as disposed to produce must themselves be represented as experiences that represent the object as red, rather than some other colour – lest the object be represented as disposed to appear something other than red. Yet these experiences can be represented as representing the object as red only if they are represented as representing it as disposed to produce experiences that represent it as red. And here the circle gets vicious. In order for an object to appear red rather than blue, it must appear disposed to appear red, rather than disposed to appear blue; and in order to appear disposed to appear red, rather than disposed to appear blue, it must appear disposed to appear disposed to appear red, rather than disposed to appear disposed to appear blue; and so on. Until this regress reaches and end, the object's appearance will not amount to the appearance of one colour rather than another. Unfortunately, the regress never reaches and end." (Boghossian & Velleman, 1989, p. 90)

This argument against dispositionalism has been criticized – and plausible ways out have been pointed at – by many authors (see (Byrne & Hilbert, 2011), (García-Carpintero, 2001) and (García-Carpintero, 2014)). Rather than delving into this discussion, though, I want to offer a reconstruction of this argument that would target CACPs such as C1 and see how its blow can be dodged. The strategy I'll suggest is, I take it, different from the one adopted by the defenders of color dispositionalism.

Consider the following schematic CACP:

(C) the Xs cooperate if and only if (a) each X intends to cooperate and (b) BLA.

And express it as an identity statement:

(C-id) cooperation is the phenomenon which occurs between some Xs if and only if (a) each X intends to cooperate and (b) BLA.

Adapted to the present discussion, Boghossian's and Velleman's argument concludes that CACPs such as (C) and (C-id) are incoherent because they make it impossible for someone to intend to cooperate. The reasoning runs as follows:

According to (C-id),

(1) x intends to cooperate *only if* x intends to [engage in the phenomenon which occurs between some ys if and only if (a) each y intends to cooperate and (b) BLA]

(2) x intends to [engage in the phenomenon which occurs between some ys if and only if (a) each y intends to cooperate and (b) BLA] *only if* x intends to [engage in the phenomenon which occurs between some ys if and only if (a) each y intends to [engage in the phenomenon which occurs between some zs if and only if (a) each z intends to cooperate and (b) BLA] and (b) BLA]

(3) …. [106]

Hence, x intends to cooperate only if this regress gets to an end. And since it doesn't, x doesn't intend to cooperate. Since x was chosen arbitrarily, we conclude that, according to (C-id), none ever intends to cooperate.

A first attempt to resist this argument exactly mimics the reply of the defenders of color dispositionalism (see (Byrne & Hilbert, 2011) and (García-Carpintero, 2014)). Following this line thought, the argument equivocates between *de re* and *de dicto* readings of the attitudes reports in (1),

[106] I use brackets to indicate the scope of each occurrence of "intend" only to improve readability; for the same reason, I italicize "only if".

(2), (3), etc. On the one hand, a *de re* reading seems to secure that each step in the regress is indeed true according to (C-id). For, assuming (C-id)'s truth, the clauses in brackets all have the same reference. But, on such a reading, there is nothing problematic about the regress. It only arises because there are infinitely many descriptions which denote cooperation – something we don't need to worry about[107] – and we should certainly not require that it comes to an end for it to be the case that someone intends to cooperate. On the other hand, while a *de dicto* reading may offer some support to the claim that the regress is vicious, it threatens to block the regress from even getting started. For it is relatively uncontroversial that *de dicto* readings of attitude verbs introduce intensional contexts in which substitution by co-referring terms fail – and so it would appear that the inferences in (1), (2), (3), etc.. are invalid.

But I'm not quite sure that friends of CACPs can get off the hook that easily. Presumably (C-id) shouldn't be seen as any old kind of identity statement – but rather as an a priori or conceptual truth. And while it is relatively uncontroversial that a posteriori identities do not support substitution in intensional contexts, the situation may be trickier when it comes to a priori or conceptual truths. Hence, some may be tempted to accept that, as a matter of necessity, if John believes that Hannah is married, then John believes that Hannah has a spouse. Of course, the view that some kind of conceptual truths support substitution in intensional contexts is itself controverted – and there would anyway be room to argue that (C-id) shouldn't be intended as *this* kind of conceptual truth (for not all conceptual truths seem to support inferences like the above: for instance, from the fact that I believe that I'm sitting it doesn't intuitively follow that I believe that [I'm sitting and that nothing is both blue and red all

[107] If you do worry, consider the following trivial identity statement, "Marie Curie is the woman who is identical to Marie Curie", from which it follows that "Marie Curies is the woman who is identical to the woman who is identical to Marie Curie", etc. hence generating an endless series of descriptions that all refer to Marie Curie.

over][108]). But these are very subtle, complex and controverted issues – and I think we'd better look for an alternative way to rescue CACPs from Boghossian's and Velleman's argument.[109]

My strategy will be to insist that, once we understand CACPs properly, we see that – no matter whether we understand the attitude reports *de re* or *de dicto* – Boghossian's and Velleman's regress isn't problematic.

Roughly put, my point is the following: Boghossian's and Velleman's regress is vicious (i.e. needs to come to an end) only on the assumption that biconditional such as (C) and identity claims such as (C-id) *exhaust the content of the notion of cooperation*. In other words, their argument crucially relies on the premise that CACPs give complete definitions of collective notions. With this assumption in place I do believe that their argument is devastating. For then, the content of an intention to cooperate really *just is* to engage in the phenomenon which crucially involves intentions to cooperate; and the content of this latter clause in turn *just is* to engage in the phenomenon which crucially involves intentions to engage in the phenomenon which crucially involves intentions to cooperate, etc. And the fact that this series doesn't come to an end does show that – under the assumption that (C) exhausts the content of the notion of cooperation – intentions to cooperate have no distinctive content.

But now relax the assumption – that is, accept that there is more to the content of the notion of cooperation than what is made explicit in (C) and (C-id). Then the content of an intention to cooperate *is not just* to engage in the phenomenon which crucially involves intentions to cooperate; and, in particular, the distinctiveness of the former doesn't depend on the distinctiveness of the latter, but rather the opposite. That is, the phrase "the phenomenon which crucially involves intentions to cooperate" has the distinctive sense it has because "cooperation" has the distinctive content it has,

[108] Once again, I use the brackets to make clear what falls within the scope of the intentional verb.
[109] This is not to say that the defenders of dispositionalism are wrong in offering the reply they offer. They may have good reason to think that color dispositionalism isn't to be understood as a conceptual thesis (although see (Lewis, 1997)) or that it isn't the kind of conceptual thesis which supports inference in intensional contexts.

not the other way around. Then, the regress (if it indeed gets off) is grounded at each stage and need not come to an end.

Let me explicate the same point from a different angle: in order to resist Boghossian's and Velleman's argument, we need to grant that CACPs only provide *parasitic* modes of presentation of the phenomenon they target, where a mode of presentation of P is parasitic if it succeeds in representing P partly in virtue of there being another (call it *independent*) mode of presentation of P. Circular identities such as (C-id) do indeed provide with a mode of presentation of the phenomenon cooperation ("the phenomenon which occurs between some Xs if and only if (a) each X intends to cooperate and (b) BLA"). But it is parasitic on another independent mode of presentation of this phenomenon – namely the concept of cooperation. And although the parasitic mode of presentation may unfold some aspects of the content of intentional states involving the concept of cooperation, such states get their content fully settled only thanks to the concept of cooperation itself. Thus, the regress need not bother us any longer. It may or may not follow from (C-id) that if I intend to cooperate, I intend to engage in the phenomenon which crucially involves intentions to cooperate. But, in any case, the content of the latter intention is settled in virtue of the settledness of the content of the former – and not the other way around. Hence, once again, the regress (if it indeed gets off) is grounded at each stage and need not come to an end.

That CACPs fail to make explicit the full content of collective notions is something that we granted from the beginning (see Section 4.2). It therefore appears that Boghossian's and Velleman's argument give us no new reason to be dissatisfied with CACPs. At any rate, it doesn't seem to give us any reason to think that they could never be true.

I grant that this reply of mine strongly invites the following questions: what is it that CACPs are missing? And aren't accounts which fully unfold collective notions way superior to CACPs? These are pressing issues indeed – issues that anyone who would want to put forward a CACP as the ultimate account of

a collective phenomenon should deal with. But I need not do so here. For I'm only claiming that CACPs can be insightful – not that they are the insuperable.[110]

4.3.2. Individuating collective phenomena

One of the main goals of the literature on collective phenomena is to pin down the hallmark of these phenomena; that is, to make explicit what distinguishes them from everything else and, in particular, what distinguishes them from non-collective sums of individual actions. Hence, typically, inquiries into the nature of collective action are introduced by comparing a non-collective aggregate of individual actions (e.g. two people who happen to be walking alongside) with a collective action (e.g. two people going for a walk together); we are then invited to consider what account for the difference in 'collectiveness' between these two scenarios (see for instance (Gilbert, 1990, pp. 2–3)).

Now, one may worry that CACPs individuate their target in a merely trivial and uninteresting way, i.e. that they only provide with trivial answers to questions such as "what makes for the difference between cooperation and non-collective sums of individual actions?", or "what is the distinctive feature of cooperative behavior?". To get a better grip on this worry, it may be worth going through an example. Consider the following circular accounts of, respectively, cities and villages:

> CITY: x is a city if and only if x involves human beings who constitute a city.
>
> VILLAGE: x is a village if and only if x involves human beings who constitute a village.

Now ask what makes for the difference between a city and a village. If we restrict ourselves to the information that CITY and VILLAGE reveal, we can only say that, while the human beings involved in a city constitute a city, those involved in a village constitute a village. But this entails that cities and villages are distinct only on the presupposition that….cities and villages are distinct. Hence, it is a trivial explanation, which presupposes what it is meant to explain.

[110] I shall nevertheless offer some thoughts that will gesture at the way in which the worries behind these questions could be answered in Section 5.

From there, it is a short leap to think that CACPs in general will exhibit the same difficulties in providing with non-trivial individuation conditions. Consider for instance the following toy circular accounts of cooperation (a collective phenomenon) and strategic interaction (a phenomenon that is usually considered as non-genuinely collective).

(C2) Agents Xs cooperate if and only if:

(a) each X intends to cooperate with the Xs

(b) each X acts in accordance with (a)

(c) (i) each X believes that (a) and (b) obtain, (ii) each X believes that (i), etc.

(SI) Agents Xs strategically interact if and only if:

(a) each X intends to strategically interact with the Xs

(b) each X acts in accordance with (a)

(c) (i) each X believes that (a) and (b) obtain, (ii) each X believes that (i), etc.

Now, by the lights of C2 and SI, cooperation and strategic interaction are different because the former involves intentions to cooperate, while the latter involves intentions to strategically interact. And this, one may worry is indeed a trivial explanation. For an intention to cooperate just is an intention of the kind involved in cooperation; and an intention to strategically interact just is an intention of the kind involved in strategic interaction. And then an intention to cooperate differ from an intention to strategically interact only if the corresponding phenomena are distinct – and hence our explanation presupposes what it is meant to explain, thus failing on the ground of triviality.

I agree with my opponent that, by the lights of C2 and SI, cooperation and strategic interaction differ because the former involves intentions to cooperate, while the latter involves intentions to strategically interact. I don't think though that this explanation is trivial. The crucial point is that an intention to cooperate is not just an intention of the kind involved in cooperation (whatever that means). Rather, it is also an intention which somehow contains the concept of cooperation, and uses it to represent a certain state of affairs (obviously, the same is true, mutatis mutandis, for an intention

to strategically interact). Hence, an intention to cooperate and an intention to strategically interact are distinct if the notions of cooperation and that of strategic interaction are distinct. And this, in turn, doesn't appear to presuppose that cooperation and strategic interaction are two distinct phenomena. After all, I can know that the notions of H_2O and that of water are distinct without knowing that they denote two different phenomena. And, accordingly, an intention to drink H_2O differs from an intention to drink water, irrespective of whether water is really H_2O.

Summing up, C2 and SI yield an explanation of the difference between cooperation and strategic interaction: these two phenomena are distinct because (1) the notions of cooperation and that of strategic interaction are distinct and (2) these notions are constitutive of the phenomena they refer to. And this explanation is non-trivial for neither does it merely restate the explanandum, nor does it contain a clause which presupposes it.

It therefore appears to me that CACPs can individuate non-trivially the phenomena they target. They can tell us that the distinctive feature of these phenomena lies in the intentionality of the agents involved. They can describe this distinctive feature in a language we understand. And they can do all that without viciously presupposing the distinctiveness of the phenomena under analysis.

4.4. Searle

In Section 4.3, I claimed that CACPs shouldn't be too swiftly dismissed, because – even though they fail in other respects – they may give interesting insights into the nature of collective phenomena. In particular, I gave an example of CACP which, if true, would strongly support the claim that the kind of intentions involved in cooperation are (regular, common garden) individual intentions with a characteristic content.

These are, I believe, important remarks that many people tend to overlook. In particular, Searle seems to consider that circularity is reason enough to fully dismiss an account, hence ignoring its implications.

And this mistake may be partly responsible for his extravagant and obscure conception of collective intention.[111]

According to Searle:

> "we-intentions are a primitive form of intentionality, not reducible to I-intentions plus mutual beliefs (…) [they] are intentions whose form is: We intend that we perform act A (…) [and postulating such intentions] only requires us to postulate that mental states can make reference to collectives where the reference to the collective lies outside the bracket that specifies the propositional content of the intentional state. The thought in the agent's mind is simply of the form "we are doing so and so." (Searle, 1990, pp. 407–408)

It is not the place here to try and explicate how we should understand the mysterious conception collective intention that Searle favors. By the light of the previous quote, one thing is clear though: he holds that whatever distinguishes collective intentions isn't their content, but rather something else – which he calls their form. I shall now argue that his argument to this claim doesn't get off the ground once we recognize that circular accounts can be valuable as sources of insights into the nature of the phenomenon they target – even though they fail in other respects, such as reducing and individuating collective notions.

Searle's reasoning to this conclusion was first laid out in his *Collective Intentions and Actions*. I reproduce the relevant passages:

> "I think most philosophers would agree that collective behavior is a genuine phenomenon, the disagreement comes in how to analyze it…Most empirically minded philosophers think that such phenomena must reduce to individual intentionality (…) I have never seen any such analysis that wasn't subject to obvious counterexamples (…) To have an actual sample analysis to work with, let us try that of Tuomela and Miller, (…) which is the best I have seen.
> Leaving out various technical details, we can summarize their account as follows: an agent A, who is a member of a group, "we-intends" to do X iff:
> 1. A intends to do his part of X.
> 2. A believes that the preconditions of success obtain, especially he believes that the other members of the group will (or at least probably will) do their parts of X.
> 3. A believes that there is a mutual belief among the members of the group to the effect that the preconditions of success mentioned in 2 above obtain.
> (…) I think it is easy to see what is wrong with the Tuomela-Miller account: A member of a group can satisfy these conditions and still not have a we-intention. Consider the following:
> Suppose a group of businessmen are all educated at a business school where they learn Adam Smith's theory of the hidden hand. Each comes to believe that he can best help humanity by pursing his own selfish interest and they each form a separate intention to this effect, i.e. each has an intention he would express as, "I intend to do my part toward helping humanity by pursuing my own selfish interest and not cooperating with anybody". Let us also suppose that the members of the group have a mutual belief to the effect that each intends to help humanity by pursuing his own selfish interests, and that these intentions will probably be

[111] In Searle's terminology, collective intentions are the mental states individual agents hold when they engage in collective action. Hence, a Searlian collective intention is a mental state of an individual (see Searle's *Constraint 1* on accounts of collective intentions (Searle, 1990, p. 406)).

> carried out with success. That is, we may suppose that each is so well indoctrinated by the business school, that each believes that their selfish efforts will be successful in helping humanity.
> Now consider any given member A of the business school graduating class.
> 1. A intends to pursue his own selfish interests without reference to anybody else, and thus, he intends to do his part toward helping humanity.
> 2. A believes that the preconditions of success obtain. In particular, he believes that other members of his graduating class will also pursue their own selfish interests, and thus help humanity.
> 3. As A knows that his classmates were educated in the same selfish ideology that he was, he believes that there is a mutual belief among the members of his group that each will pursue their own selfish interests, and that this will benefit humanity.
> Thus, A satisfies the Tuomela-Miller conditions, but all the same, he has no collective intentionality. There is no we-intention. There is even an ideology, which he and the others accept, to the effect that there should not be a "we-intention".
> (…)
> Could we avoid such counterexamples by construing the notion of "doing his part" in such a way as to block them? I think not. We are tempted to construe "doing his part" to mean doing his part toward achieving the collective goal. But if we adopt that move, then we have included the notion of a collective intention in the notion of "doing his part". We are thus faced with a dilemma: if we include the notion of collective intention in the notion of "doing his part", the analysis fails because of circularity; we would now be defining we-intentions in terms of we-intentions. If we don't so construe "doing his part", then the analysis fails because of inadequacy. Unless the "we-intention" is built into the notion of "doing his part", we will be able to produce counterexamples of the sort I have outlined above.
> I have not demonstrated that no such analysis could ever succeed. I am not attempting to prove a universal negative. But the fact that the attempts that I have seen to provide a reductive analysis of collective intentionality fail for similar reasons—namely, they do not provide sufficient conditions for cooperation; one can satisfy the conditions in the analysis without having collective intentionality—does suggest that our intuition is right: we-intentions are a primitive phenomenon." (Searle, 1990, pp. 404–405)

Searle's argument can be summarized as follow:

> Unless they are interpreted in a way that makes them circular, all the accounts of collective intentions in terms of individual intentionality that have been offered so far have counterexamples.
>
> This gives us defeasible reason to believe that collective intentions cannot be reductively analyzed in terms of individual intentionality.
>
> ---
>
> We thus have defeasible reason to believe that collective intentions are irreducible to individual intentionality and in particular that they aren't individual intentions with a distinctive content (in Searle's terminology, they aren't I-intentions with a distinctive content).

As far as I know, this argument has been resisted either by rejecting the first premise (for instance by Bratman (Bratman, 1993)), or by simply insisting on how mysterious and uninformative is the view that it led Searle to adopt (see (Gold & Sugden, 2007, p. 114)). My point is a different one: the argument is, I believe, a *non-sequitur*. Even if we accept both premises, we have no reason to accept its conclusion.

For, as I've been arguing in Section 4.3.1, a CACP can strongly support the claim that collective intention (the mental states of the agents who engage in collective behavior) are common garden individual intentions with a distinctive content (in Searle's terminology, I-intentions with a distinctive content).

Curiously enough, this is exactly the case of Tuomela's and Miller's account – interpreted in the way which, according to Searle, allows avoiding counterexamples at the cost of going circular. Thus interpreted, the account reads:

(TMS[112]) An agent "we-intends" to do X if and only if:

1. A intends to do her part towards achieving the collective goal X
2. A believes that the preconditions of success obtain, especially she believes that the other members of the group will (or at least probably will) do their parts of X.
3. A believes that there is a mutual belief among the members of the group to the effect that the preconditions of success mentioned in 2 above obtain.

I don't find it obvious that we avoid Searle's counterexample by construing "doing her part" in this way. This is so because the phrase "collective goal" is ambiguous. On a natural reading, it means a goal that several people have – and so interpreted, it doesn't suffice to rule out Searle's counterexample. Some (like, apparently, Searle) may think that the expression has a more stringent interpretation, one that makes it impossible for someone to intend to do her part of a collective goal G and not to intend to cooperate towards G. Such a reading would indeed rule out Searle's counterexample… but at least *I* have trouble grasping it. This needs not worry us though. For Searle himself appears to grant that TMS would resist his counterexample – which puts us in a dialectically safe position. Furthermore, it doesn't take much effort to find a (presumably circular) construal of "doing her part" which clearly rules out Searle's scenario. My personal favorite lays in the vicinity of "doing her part in the cooperative

[112] For Tuomela and Miller by Searle.

effort towards X".[113] I shall use TMS* to refer to the account we get when we substitute "A intends to do her part in the cooperative effort towards X" for clause 1 in TMS.

Hence, be it TMS or TMS* or both, there is a circular implementation of Tuomela's and Miller's account of we-intention that avoids Searle's counterexample. Hence, it seems contextually appropriate to consider it as true. But now, crucially, the truth of either TMS or TMS* would strongly support the view that, contra Searle, we-intentions are I-intentions with a distinctive content. For, according to these accounts, it is both sufficient and necessary that an agent has some I-intentions and some I-beliefs, for her to have a we-intention. Hence, without further reason to do so, postulating that we-intentionality is something over and above common garden individual intentionality seems clearly unwarranted.

As I already said, I believe that Searle's mistake is to think that circularity suffices to fully dismiss an account, hence disregarding its implications. As I've already granted in Section 4.2.1, he is right in considering that accounts such as TMS and TMS* aren't *reductive analyses* of the notion of we-intention – and this entails that they fail to show how to reduce the notion of we-intention to other, non-collective, concepts. But it crucially doesn't follow that such accounts can be dismissed and ignored. For, their circularity notwithstanding, they can still reveal important features of the phenomenon they target – and, in particular, they may still support a reductive identification (p.125) of this phenomenon.

Searle could try replying, I believe, in two different ways. Firstly, he might claim that collective notions such as cooperation could never figure in the content of a common garden I-intention. But, short of an argument for this very surprising claim (after all, it surely seems to me that I intend to do my bit in a cooperative effort towards a less sexist world), this option is a no-go. And the argument is missing. Secondly, he may insist that accounts such as TMS and TMS* cannot be true, because they are inconsistent. But, as already claimed in Section 4.3.1, there is nothing obviously inconsistent in CACPs,

[113] I take it to be fairly obvious that in Searle's example, the businessmen don't intend to do their part of a cooperative effort towards helping humanity. Rather, in order to achieve this goal, they intend to avoid cooperation by all means.

and the most prominent argument trying to establish this conclusion fails, once CACPs are understood with the proper restriction.

4.5. Conclusion: beyond CACPs?

In this chapter, I've been discussing the value of circular accounts of collective phenomena. I've granted that CACPs aren't reductive analyses and that they fail to fully unfold the content of collective notions (Section 4.2). I've also argued that, this notwithstanding, CACPs may be valuable as sources of insights and ways of individuating the phenomena they target (Section 4.3). Furthermore, I've shown that ignoring the insights that CACPs contain may lead someone astray (Section 4.4). I therefore conclude that, in spite of their limitations, CACPs shouldn't be swiftly dismissed, and that we should take care not to ignore their implications.

As a manner of conclusion, I want now to briefly consider a question that has surfaced in several places in this chapter. Could CACPs be the best accounts of collective phenomena we can come up with? Or should we rather always expect that there are alternative superior options (maybe yet to be discovered)? It isn't the place to argue for a definitive answer, but here come a few thoughts.

Suppose that a given CACP, whose target is cooperation, is true – and call it C. This is, I believe, compatible with two very different types of situations. On the one hand, cooperation may be liable to reductive analysis. On the other, cooperation may be a basic notion.

In the first situation, there is a reductive analysis of cooperation (may be yet to be discovered) which entails C's truth.[114] Such a reductive analysis is properly superior to C in every respect. Since it entails C's truth, it contains all the insights C contains. Furthermore, since it is a reductive analysis, it shall fully unfold the content of the notion of cooperation, henceforth proving satisfactory where C was dissatisfying. Thus, if collective notions are reducible, CACPs are nothing more than valuable steps

[114] Given the comments of Chapter 2, Section 3.3. (p.63) on the recursive nature of overt belief and common knowledge (and, presumably, other specifications of the notion of mutual awareness), it is to be expected that such accounts will have built into them some kind of mutual awareness condition.

towards the discovery of the ultimate non-circular accounts of collective phenomena (notice that everything I've said in Sections 4.2 to 4.4 is compatible with this scenario).

But if, as per the second type of situation, cooperation is a basic notion, it is at best unclear that we can find an account which avoids C's flaws while preserving its virtues. Suppose for instance that some higher mammals are biologically predisposed to acquire the notion of cooperation when exposed to instances thereof. In other words, suppose that we have an innate predisposition to distinguish cooperative from non-cooperative behavior – and hence intentions to cooperate from non-cooperative intentions. In such circumstances, it isn't obvious that we should be able to capture the distinction that this innate predisposition of ours tracks in *a more informative way than the one provided by C and its likes.*

Now, as far as I know, we shouldn't assume that collective notions are reducible at the outset of an investigation into the nature of collective phenomena. Rather, claiming that collective notions are not primitive is something we may want to do after carrying out such an investigation, and only if this investigation has yield (something close enough to) reductive analyses of collective notions. Hence until reductive analyses of collective notions have been provided, we should consider it an open possibility that CACPs give us the best accounts of collective phenomena we can come up with.

In this thesis, I've offered analyses of several collective phenomena (mainly cooperation in Chapter 2, and institutional rules in Chapter 3). As argued at some length, these analyses are least not apparently circular. But there is certainly some room to argue that, in ultimate instance, notions I've appealing to (such as, in Chapter 2, the notions of common goal and shared plan; and, in Chapter 3, the notion of commitment) do introduce circularity in my proposals. Albeit this would be an unwelcome result, the discussion I've had in this chapter secures that, even in these circumstances, my accounts could be insightful and, in particular, that they could nevertheless serve the overall purpose of this thesis and help us writing social reality into the book of the world – that is, help us understand how social reality emerges from presumably more fundamental aspects of reality.

The main conclusions of this thesis are:

- Social phenomena are those aspects of reality which essentially involve cooperation (Chapter 1)
- Cooperation is the phenomenon which takes place when several agents pursue a goal they have in common by following a plan they share in a state of mutual awareness. (Chapter 2)
- Humans create institutions by engaging in cooperative activities that publicly commit them to certain rules, which they have the power to enforce. (Chapter 3)

Together, these conclusions allow building a partial description of social reality which reveals how the social phenomena it accounts for relate to other, plausibly more fundamental, aspects of reality. More specifically: Chapter 2 makes explicit how cooperation, the social atom, emerges out of interrelated individual actions and intentions and Chapter 3 explicates how power and cooperation can give rise to institutional phenomena. The main purpose of this thesis – i.e. writing social reality into the book of the world – thus appears to be fulfilled to a reasonable extent.

Furthermore, beside these three conclusions concerning the nature of social reality, I have argued in Chapter 4 that circular accounts of collective phenomena can be illuminating in many ways and, in particular, can support reductive identifications of such phenomena.

As a matter of concluding thoughts, I shall now point at several routes of researches on which the present investigation may be continued:

1. Most obviously, the present investigation may be extended by building a reductive description of social reality that would cover more ground. These further investigations shall be guided by the conclusion of Chapter 1 – which implies that, one way or another, cooperation will be involved in the

emergence of every social phenomena. Among the most conspicuous social phenomena that were left out in this thesis, I reckon:

- Social groups, with respect to which the conclusion of Chapter 1 suggest the following questions: how does cooperation contribute to the existence of such groups? And if, as it seems plausible, there is no general answer to this question: can we build an informative taxonomy of social groups on the basis of the way how they depend upon cooperation for their existence?
- Unintended social phenomena (such as inflation, wealth inequalities, gender discrimination), with respect to which the conclusion of Chapter suggests, once again, the following question: how does cooperation contribute to the existence of such phenomena?

2. It tends to be a common assumption that institutional rules must be collectively known in a sense that imply that, if R is an institutional rule of an anchor A, at least one person involved in A must be aware of that (see (Searle, 2009, pp. 116–120) & (Thomasson, 2003b)). But now, an interesting feature of the account of institutional rules I've offered is that, at least on the face of it, it leaves open the possibility that a rule R could be the rule of an institutional anchor A even though none of the people involved in A know that. For, as argued in Section 3.4.2, one can be committed to a rule without knowing the content of this rule. So could it be the case that, after all, we can discover, in the fuller sense of the term, an institutional rule of ours? This seems to be a question worth exploring.

3. In Section 2.3.3. I've said that it is presumably in the nature of common knowledge to induce some kind of self-referentiality. Admittedly, the motivation I offered for this claim wasn't completely satisfactory.[115]But now, at least on the face of it, the implications of this claim are momentous: it would yield a general and non-mysterious characterization of the so-called self-referentiality of social

[115] I overtly stated that the argument I was offering was rough, and referred to Lewis's and Gilbert's work on the topic for further development.

concepts (see (Searle, 1995, p. 32)), one of the most intriguing feature of social reality. Therefore, this very particular aspect of my dissertation seems worth further developments.

4. In my discussion of institutional reality, I rely on the – unargued for – assumption that "if we can account for the fact that a given society has the institutional rules it has, then we have a full account of that society's institutional reality" (p.90). Exploring this assumption, putting it into question and – maybe – defending shall constitute a fruitful way to complement the work I've done so far.

References

Alonso, F. M. (2009). Shared Intention, Reliance, and Interpersonal Obligations. *Ethics*, *119*(3), 444–475.

Andersson, Å. (2007). *Power and Social Ontology* (dissertation). Lund University. Retrieved from http://lup.lub.lu.se/record/27182

Audi, P. (2012). Grounding: Toward a Theory of the In-Virtue-of Relation. *Journal of Philosophy*, *109*(12), 685–711.

Baier, A. (1986). Trust and Antitrust. *Ethics*, *96*(2), 231–260.

Bardsley, N. (2006). On collective intentions: collective action in economics and philosophy. *Synthese*, *157*(2), 141–159. https://doi.org/10.1007/s11229-006-9034-z

Blomberg, O. (2013). *Joint Action Without and Beyond Planning*. PhD Thesis.

Blomberg, O. (2015). Common Knowledge and Reductionism about Shared Agency. *Australasian Journal of Philosophy*, 1–12. https://doi.org/10.1080/00048402.2015.1055581

Blomberg, O. (2016). Shared Intention and the Doxastic Single End Condition. *Philosophical Studies*, *173*(2), 351–372.

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, *77*(1), B25–B31. https://doi.org/10.1016/S0010-0277(00)00096-2

Boghossian, P. A., & Velleman, J. D. (1989). Color as a Secondary Quality. *Mind*, *98*(January), 81–103.

Bratman, M. (1999). I Intend That We J. In *Faces of Intention: Selected Essays on Intention and Agency* (pp. 142–161). Cambridge University Press.

Bratman, M. (2009). Modest Sociality and the Distinctiveness of Intention. *Philosophical Studies*, *144*(1), 149–165.

Bratman, M. E. (1992). Shared Cooperative Activity. *Philosophical Review*, *101*(2), 327–341.

Bratman, M. E. (1993). Shared Intention. *Ethics*, *104*(1), 97–113.

Bratman, M. E. (2014a). Rational and Social Agency: Reflections and Replies. In M. Vargas & G. Yaffe (Eds.), *Rational and Social Agency* (pp. 294–344). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199794515.003.0012

Bratman, M. E. (2014b). *Shared Agency: A Planning Theory of Acting Together* (1 edition). New York, NY: Oxford University Press.

Burgess, J. a. (2008). When Is Circularity in Definitions Benign? *The Philosophical Quarterly*, *58*(231), 214–233. https://doi.org/10.1111/j.1467-9213.2007.522.x

Butterfill, S. (2012). Joint Action and Development. *The Philosophical Quarterly*, *62*(246), 23–47. https://doi.org/10.1111/j.1467-9213.2011.00005.x

Byrne, A., & Hilbert, D. R. (2011). Are Colors Secondary Qualities? In L. Nolan (Ed.), *Primary and Secondary Qualities: The Historical and Ongoing Debate*. Oxford University Press.

Call, J., & Tomasello, M. (2008). Does the Chimpanzee Have a Theory of Mind? 30 Years Later. *Trends in Cognitive Sciences*, *12*(5), 187–192.

Carpenter, M., & Call, J. (2013). How Joint Is the Joint Attention of Apes and Human Infants? In J. Metcalfe & H. S. Terrace (Eds.), *Agency and Joint Attention* (pp. 49–61). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199988341.003.0003

Chant, S. R. (2006). The Special Composition Question in Action. *Pacific Philosophical Quarterly*, *87*(4), 422–441.

Chant, S. R. (2007). Unintentional Collective Action. *Philosophical Explorations*, *10*(3), 245–256.

Correia, F., & Schnieder, B. (2012). *Metaphysical Grounding: Understanding the Structure of Reality*. Cambridge University Press.

Crespi, B. J., & Yanega, D. (1995). The definition of eusociality. *Behavioral Ecology*, *6*(1), 109–115. https://doi.org/10.1093/beheco/6.1.109

Daly, C. (2012). Scepticism About Grounding. In *Metaphysical Grounding: Understanding the Structure of Reality* (p. 81). Cambridge University Press.

Davies, M. (1987). Relevance and Mutual Knowledge. *Behavioral and Brain Sciences*, *10*(4), 716.

Diaz-Leon, E. (2013). What Is Social Construction? *European Journal of Philosophy*, n/a-n/a. https://doi.org/10.1111/ejop.12033

Dummett, M. (1975). Wang's Paradox. *Synthese*, *30*(3–4), 201–32.

Dworkin, R. (1977). *Taking Rights Seriously*. Cambridge, Mass: Harvard University Press.

Dworkin, R. M. (2013). *Justice for Hedgehogs* (Edición: Reprint). Cambridge, Mass.: The Belknap Press.

Engel, P. (1998). Believing, Holding True, and Accepting. *Philosophical Explorations*, *1*(2), 140–151.

Epstein, B. (2009). Ontological Individualism Reconsidered. *Synthese*, *166*(1), 187–213.

Epstein, B. (2014). Social Objects without Intentions. In A. K. Ziv & H. B. Schmid (Eds.), *Institutions, Emotions, and Group Agents* (pp. 53–68). Springer Netherlands. Retrieved from http://link.springer.com/chapter/10.1007/978-94-007-6934-2_4

Fine, K. (1994). Essence and Modality. *Philosophical Perspectives*, *8*, 1–16.

Fine, K. (2012). Guide to Ground. In F. Correia & B. Schnieder (Eds.), *Metaphysical Grounding* (pp. 37–80). Cambridge University Press.

García-Carpintero, M. (2001). SENSE DATA: THE SENSIBLE APPROACh. *Grazer Philosophische Studien*, *62*(1), 17–63.

García-Carpintero, M. (2014). Josep Corbí, Morality, SelfKnowledge and Human Suffering: An Essay on the Loss of Confidence in the World, London: Routledge, 2012, Xvi + 254 Pp. GBP 80.00 (Hardback), ISBN 9780415890694. *Dialectica*, *68*(1), 151–161.

Gilbert, M. (1978). *On Social Facts*. St. Hilda's College, Oxford.

Gilbert, M. (1992). *On Social Facts*. Princeton University Press.

Gilbert, M. (1998). In Search of Sociality. *Philosophical Explorations*, *1*(3), 233–241. https://doi.org/10.1080/10001998098538702

Gilbert, M. (2003). The structure of social atom: Joint Commitment as the Foundation of Human Social Behavior. In F. Schmitt (Ed.), *Socializing Metaphysics: The Nature of Social Reality* (Rowman & Littlefiled Publishers, pp. 39–65). Oxford.

Gilbert, M. (2013). *Joint Commitment: How We Make the Social World*. OUP USA.

Gilbert, M. P. (1990). Walking Together: A Paradigmatic Social Phenomenon. *Midwest Studies in Philosophy*, *15*(1), 1–14.

Gold, N., & Sugden, R. (2007). Collective Intentions and Team Agency. *Journal of Philosophy*, *104*(3), 109–137.

Gopnik, A., & Slaughter, V. (1991). Young Children's Understanding of Changes in Their Mental States. *Child Development*, *62*(1), 98–110. https://doi.org/10.1111/j.1467-8624.1991.tb01517.x

Hacking, I. (1999). *The Social Construction of What?* Harvard University Press.

Haslanger, S. (1995). Ontology and Social Construction. *Philosophical Topics*, *23*(2), 95–125.

Haslanger, S. (2012). Social Construction: The "Debunking" Project. In *Resisting Reality: Social Construction and Social Critique* (pp. 113–138). Oxford: Oxford University Press.

Higginbotham, J. (2006). Sententialism: The Thesis That Complement Clauses Refer to Themselves. *Philosophical Issues*, *16*(1), 101–119. https://doi.org/10.1111/j.1533-6077.2006.00105.x

Hindriks, F. (2009). Constitutive Rules, Language, and Ontology. *Erkenntnis*, *71*(2), 253–275.

Hobbes, T. (1982). *Leviathan* (59364th edition). Harmondsworth: Penguin Classics.

Humberstone, I. L. (1997). Two Types of Circularity. *Philosophy and Phenomenological Research*, *57*(2), 249–280. https://doi.org/10.2307/2953718

Jenkins, C. S. (2011). Is Metaphysical Dependence Irreflexive? *The Monist*, *94*(2), 267–276.

Keefe, R. (2002). When Does Circularity Matter? *Proceedings of the Aristotelian Society*, *102*, 275–292.

Kutz, C. (2000). Acting Together. *Philosophy and Phenomenological Research*, *61*(1), 1–31.

Lewis, D. (1969). *Convention: A Philosophical Study*. Harvard University Press.

Lewis, D. (1978). Truth in Fiction. *American Philosophical Quarterly*, *15*(1), 37–46.

Lewis, D. (1997). Naming the Colours. *Australasian Journal of Philosophy*, *75*(3), 325–42.

Locke, J. (1993). *Two Treatises of Government* (Reprint edition). London: Everyman Paperback.

Ludwig, K. (2007). Collective Intentional Behavior From the Standpoint of Semantics. *Noûs*, *41*(3), 355–393.

Mason, R. (2016). The Metaphysics of Social Kinds. *Philosophy Compass*, *11*(12), 841–850.

Moses, L. J. (1993). Young children's understanding of belief constraints on intention. *Cognitive Development*, *8*(1), 1–25. https://doi.org/10.1016/0885-2014(93)90002-M

Newlands, S. (2013). Spinoza's Modal Metaphysics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2013). Retrieved from http://plato.stanford.edu/archives/win2013/entries/spinoza-modal/

Peacocke, C. (2005). Joint Attention: Its Nature, Reflexivity, and Relation to Common Knowledge. In N. M. Eilan, C. Hoerl, T. McCormack, & J. Roessler (Eds.), *Joint Attention: Communication and Other Minds* (p. 298). Oxford University Press.

Petersson, B. (2007). Collectivity and Circularity. *Journal of Philosophy*, *104*(3), 138–156.

Rawls, J. (1999). *A Theory of Justice* (Edición: Rev ed.). Cambridge, Mass: Harvard University Press.

Rosen, G. (2010). Metaphysical Dependence: Grounding and Reduction. In *Modality: Metaphysics, Logic, and Epistemology* (pp. 109–36). Oxford University Press.

Rousseau, J.-J. (2014). *The Social Contract*. Place of publication not identified: CreateSpace Independent Publishing Platform.

Schaffer, J. (2009). On What Grounds What. In D. Manley, D. J. Chalmers, & R. Wasserman (Eds.), *Metametaphysics: New Essays on the Foundations of Ontology* (pp. 347–383). Oxford University Press.

Schaffer, J. (2012). Grounding, Transitivity, and Contrastivity. In F. Correia & B. Schnieder (Eds.), *Metaphysical Grounding: Understanding the Structure of Reality* (pp. 122–138). Cambridge University Press.

Schiffer, S. R. (1987). *Remnants of meaning*. MIT Press.

Searle, J. (1990). Collective Intentions and Actions. In *Intentions in Communication* (pp. 401–415). MIT Press.

Searle, J. (2009). *Making the Social World: The Structure of Human Civilization*. Oxford University Press.

Searle, J. R. (1995). *The Construction of Social Reality*. Simon and Schuster.

Searle, J. R. (2005). What is an institution? *Journal of Institutional Economics*, *1*(1), 1–22. https://doi.org/10.1017/S1744137405000020

Shapiro, S. J. (2014). Massively Shared Agency. In M. Vargas & G. Yaffe (Eds.), *Rational and Social Agency* (pp. 257–289). Oxford University Press. Retrieved from http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199794515.001.0001/acprof-9780199794515-chapter-11

Sider, T. (2014). *Writing the Book of the World* (Reprint edition). Oxford: Oxford University Press.

Skiles, A. (2015). Essence in Abundance. *Canadian Journal of Philosophy*, *45*(1), 100–112.

Smith, M., Lewis, D., & Johnston, M. (1989). Dispositional Theories of Value. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *63*, 89–174.

Thomasson, A. L. (2003a). Foundations for a Social Ontology. *Protosociology*, *18–19*, 269–290.

Thomasson, A. L. (2003b). Realism and Human Kinds. *Philosophy and Phenomenological Research*, *67*(3), 580–609. https://doi.org/10.1111/j.1933-1592.2003.tb00309.x

Thomasson, A. L. (2007). *Ordinary Objects*. Oxford University Press.

Thomasson, A. L. (2009). Social Entities. In R. L. Poidevin (Ed.), *The Routledge Companion to Metaphysics*. Routledge.

Tollefsen, D. (2005). Let's Pretend!: Children and Joint Action. *Philosophy of the Social Sciences*, *35*(1), 75–97.

Tomassello, M. (1995). Joint attention as social cognition. In *Joint attention: Its orgins and role in development* (Lawrence Erlbaum, pp. 103–130). Hillsdale, NJ: C.Moore and P.Dunham.

Tuomela, R. (1993). What Is Cooperation? *Erkenntnis (1975-)*, *38*(1), 87–101.

Tuomela, R. (2000). Collective and Joint Intention. *Mind and Society*, *1*(2), 39–69.

Tuomela, R. (2003). Collective Acceptance, Social Institutions, and Social Reality. *American Journal of Economics and Sociology*, *62*(1), 123–165. https://doi.org/10.1111/1536-7150.t01-1-00005

Tuomela, R. (2016). Cooperation as Joint Action. *Analyse & Kritik*, *33*(1), 65–86. https://doi.org/10.1515/auk-2011-0106

Tuomela, R., & Miller, K. (1988). We-Intentions. *Philosophical Studies*, *53*(3), 367–389.

Wellman, H. M., & Bartsch, K. (1988). Young children's reasoning about beliefs. *Cognition*, *30*(3), 239–277.

Williamson, T. (1994). *Vagueness* (Vol. 81). Routledge.

Wilson, J. M. (2014). No Work for a Theory of Grounding. *Inquiry*, *57*(5–6), 535–579.

Ylikoski, P., & Mäkelä, P. (2002). We-Attitudes and Social Institutions. In *Social Facts and Collective Intentionality*. Dr. Hänsel-Hohenhausen AG.