The University Of Sheffield.

Problematic Situations, Preference Change, and Negotiations: A Philosophical Approach

By:

David Strohmaier

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Arts and Humanities
Department of Philosophy

May 2018

Abstract:

For many of the large-scale problems facing humanity, individuals lack the power to address them on their own. We might call upon group agents such as states or corporations to solve them for us. Problems such as climate change, however, are of such a scope and magnitude that no single group agent can deal with them effectively. To achieve a cooperative solution, groups must negotiate with one another on how to address these problems. Contributing to our understanding of how humanity can deal with such large-scale problems, the present thesis offers a theory of negotiating group agents.

After an introduction, chapters two, three, and four offer a philosophical reconstruction of the sociological Negotiated Order approach. At the core of these chapters is a pragmatist theory of motivational change in social contexts, which I contrast with standard rational choice theory. The Negotiated Order approach argues that many social phenomena, such as organisations, function based on motivational change occurring in the context of informal negotiations.

Chapter five discusses group agency and argues that functionalist accounts of group agency are a promising approach for extending the Negotiated Order approach. The sixth and final chapter returns to the original motivation for developing a theory of negotiating group agents. It shows that in the case of climate change negotiations, the Negotiated Order account offers a different and promising perspective than standard rational choice models.

In addition, the thesis includes an appendix which discusses Dewey's theory of choice and proposes a way of formalising its pragmatist take on preference change.

Contents

Chapter One: Introduction

Whether it be climate change, transnational economic crises, or the threat of nuclear war, humanity is facing large-scale problems. In light of these pressing problems, we might call upon agents to work out a resolution. But individuals lack the power to effectively address these problems and humanity itself is not an agent[1]. Group agents, however, serve as good candidates for addressing them.

By forming group agents, human individuals pool their power. States, corporations, or transnational NGOs may be among the problem-solvers needed. In recent years, various philosophers (List & Pettit 2011, Huebner 2014, Tollefsen 2015, Epstein 2017) have put forward theories of group agency that allow us to see which groups might be eligible candidates for acting on climate change, economic crises, or threats of war.

The aforementioned problems, however, are of such scope and magnitude that no single group agent can deal with them effectively. For example, while we might place our hopes in the USA's contribution to combating climate change, even this world power is not up to the task on its own (see Gardiner 2011: 96-97 for an explanation as to why not). Multiple group agents must act *together* to avoid catastrophe although their individual interests will diverge. Thus to achieve a cooperative solution, they must *negotiate* with one another on how best to address the problems.

We therefore need to understand how group agents negotiate with one another. What kind of negotiators are states when they face one another in climate change negotiations? Which assumptions should we make when we model the negotiations between group agents? To answer these questions, I propose a theory of negotiating group agents.

[1] Or at least it is not based on the more plausible theories of group agency: cf. Lawford-Smith 2015.

Standard rational choice theory offers one account of negotiations. By standard rational choice theory, I am referring to accounts in decision and game theory which assume that intrinsic preferences are fixed [2] and that agents, individual or group, act on their preferences and beliefs in an attempt to maximise their preference satisfaction. Christian List and Philip Pettit put forward an account of group agents built upon standard rational choice theory in their influential book, *Group Agency* (2011).[3]

While the assumptions of standard rational choice theory allow the construction of simple and powerful mathematical models, they also limit the usefulness of the theory as an account of negotiations. From the vantage point of standard rational choice theory, some potential avenues for advancement in dealing with large-scale problems remain undetectable, or so I suggest. One might hope to find options other than the one outlined in the occasionally dismal models of rational choice theory.

For example, in climate change negotiations, states might seem caught up in a prisoner's dilemma or a tragedy of the commons (Gardiner 2011). States prefer mutual cooperation – each state curbing its emissions – over mutual defection resulting in severe climate change. But the states also prefer the option of defecting while others cooperate. In other words, each state prefers it that the others cut their emissions to avert climate change, leaving its own economy to progress unhindered. If countries also deem it preferable not to be the "sucker", (that is, not to be the one who cooperates while others withdraw,) defection dominates. No matter what the other agents do, each one maximises their satisfaction of preferences by defecting.

Being rational preference maximisers, the states defect and the climate deteriorates. But not only would the climate suffer – the overall pay-off would be even lower than with

[2] For the notion of intrinsic preferences, see Binmore 2009: 5-6

[3] Although List developed more sophisticated models of rational choice with Franz Dietrich (2013), we find little of this reflected in List's work on group agency.

mutual cooperation. We want to avoid such an outcome, but how can we do this, given the assumption of standard rational choice theory? I propose that standard rational choice theory occludes certain ways forward, and that abandoning its dubious assumptions allows us to see these options. A different view on negotiations is the first step towards avoiding what seems to be inevitable failure from the perspective of the standard approach.

The present thesis offers a different theory of negotiating group agents. Drawing on the sociological Negotiated Order[4] approach, I provide an account of group agents exhibiting a so-called Negotiated Order internally and in their interactions with one another. Within sociology, the Negotiated Order approach, spearheaded by Anselm Strauss, presents some competition to standard rational choice approaches. It claims to offer a different account of agency that incorporates motivational change in an original manner. I dedicate a large part of my thesis to reconstructing this account of agency and showing how it diverges from standard rational choice theory.

My thesis presents the account of negotiating group agents that emerges if we base our account on the Negotiated Order approach rather than standard rational choice theory. By endorsing the Negotiated Order approach, a picture of group agents results that allows us to see new options for negotiating large-scale problems. To establish this broadening of perspective, I look especially at models of climate change negotiations between states.

I illustrate how alternative ways out of unfruitful negotiation deadlocks can be discovered when states undergo a type of preference change as proposed by the Negotiated Order approach. Abandoning the assumption that intrinsic preferences exist prior to the

[4] To indicate the technical nature of this term, I will capitalise it here and in the following. The exact analysis will occur in chapter four.

situation of action, new potential for cooperation becomes apparent in our account of climate change negotiations.

My thesis does not establish the correctness of the Negotiated Order approach, which would be an empirical endeavour in sociology and psychology, nor does it defend the reality of group agents, as much of the current literature in the philosophical debate attempts to do. Rather, I reconstruct a sociological approach and demonstrate that when endorsed, the approach provides us with a different picture of the interaction between multiple group agents than thus far assumed in philosophical debates (for example by Gardiner 2011).

Before I provide an outline describing how my thesis develops the Negotiated Order approach as a theory of negotiating group agents, it is necessary to go into more detail about the sociological approach. Since the Negotiated Order approach has received little philosophical attention so far, I introduce it and its history. This background information makes it easier to see why I proceed the way I do.

The History and Context of the Negotiated Order Approach

The roots of the Negotiated Order approach reach back to the institutionalisation of sociology in the USA and, in particular, the Chicago School of the early 20th century. On the sociological side, William Thomas and Robert E. Park numbered among the most important members, but the school always exhibited a philosophical side. Not only had Park been under Dewey's influence (cf. Joas 1993: 33), but George Herbert Mead remained in close contact with the future sociological researchers.

The Chicago School did not propagate one principled sociological theory, but rather formed a loosely united community. As Paul Rock writes, "[f]from the very first, an oral

tradition dwarfed the formal and articulate processes of disseminating ideas" (Rock 1979: 5). But without doubt, classical pragmatist philosophy influenced the school and its account of individual agents from the start (Joas & Knöbl 2009: 184-185). For this reason, I occasionally call the sociological approaches following the Chicago School "pragmatist sociology".

One particular element of Dewey's theory of action lived on in pragmatist sociology: the Problematic Situation.[5] According to Dewey, Problematic Situations prompt agents to change their take on the world (cf. Joas & Knöbl 2009: 189). How exactly we should describe such Problematic Situations and what it means for agents to change their take on the world is the main topic of the next chapter.

During its heyday, the Chicago School produced prominent sociological research, especially in the field of urban sociology, but starting in the 1940s its influence waned. The heritage of pragmatist sociology and of the Chicago School continued in symbolic interactionism, which has been recognised as a distinct school of sociology ever since. Symbolic interactionism took its cues from the earlier Chicago School and the work of George Herbert Mead.[6] Herbert Blumer, who had been Mead's assistant, became the father and most important proponent of symbolic interactionism. In Blumer's work, Chicago's oral tradition lived on.

The symbolic interactionists look closely at how individuals interact, how they interpret the situation and each other during interactions, and how these interpretations shift. With the emphasis on meaning and interpretation came a narrowing of the perspective of symbolic interactionism. While Blumer continued the pragmatist tradition in sociology, some questions raised by the Chicago School received less attention from the symbolic

[5] To indicate the technical nature of this term, I will capitalise it here and in the following.
[6] For a debate within sociology on the pragmatist roots of symbolic interactionism, see the exchange between Lewis (1976, 1977) and Blumer (1977).

interactionists. The original Chicago School had been famous for its urban sociology, which focussed on the structure of cities and addressed issues at various scales. The symbolic interactionists, on the other hand, focussed on interaction in the microsphere, that is among a small number of individuals. They engaged in a form of folk psychological analysis of interacting agents interpreting various objects, including each other. The symbolic interactionists rarely considered social phenomena on a larger scale.

As a consequence, critics accuse symbolic interactionism as lacking a proper theory of sociological macro-phenomena. According to this criticism, the symbolic interactionists fail to account for large-scale phenomena. While many sociologists use insights into micro-interaction offered by symbolic interactionism, they hold that symbolic interactionism remains incomplete at best, since it neglects macro-phenomena. Climate negotiations, involving virtually all nations of the world, are one such macro-phenomenon. Although attempts have been made to respond to this accusation, the problems with macro-phenomena have lingered with symbolic interactionism right through to the present (cf. Joas & Knöbl 2009: 216-219).

Anselm Strauss put forward his Negotiated Order approach in the context of symbolic interactionism and its limitations. Strauss had received his education at Chicago University and contributed to the formation of symbolic interactionism with his early book (1997 [1959]) on identity entitled *Mirrors and Masks*. His Negotiated Order approach turned the pragmatist tradition in the form of symbolic interaction towards organisation theory. Like many authors in the tradition of the Chicago School, Strauss developed his approach by engaging in empirical research.

Together with a team, Strauss investigated medical institutions and, in 1963, published a foundational text for the Negotiated Order approach: “The Hospital and Its Negotiated Order”. This book chapter, which I draw on repeatedly, presents the main tenets of the

Negotiated Order approach. Further important sources are Strauss' books *Negotiations* (1988 [1978]), which elaborates and defends the Negotiated Order approach, and *Continual Permutations of Actions* (2014), which discusses the theory of action underlying Strauss' work.

Being an offspring of symbolic interactionism, the negotiated approach inherited the problems regarding macro-phenomena. While these issues are not the main topic of my thesis, the focus on group agents has an interesting consequence: It 'scales up' pragmatist sociology. By discussing how the Negotiated Order approach sheds light on climate negotiations between states, I show how it can address a macro-phenomenon. Bearing in mind this background knowledge of pragmatist sociology, I can clarify how my own project proceeds.

Outline of a Theory of Negotiating Group Agents

To reap the benefits of the Negotiated Order approach for a theory of negotiating group agents, I reconstruct it in a way palatable for philosophers. The pragmatist heritage of the approach serves as the starting point for this endeavour.

The second chapter of my thesis introduces Dewey's theory of the Problematic Situation mentioned above. I reconstruct this theory as an account of preference change, which forms the first pillar of my reconstruction of the Negotiated Order approach. Since my reconstruction of the Negotiated Order approach serves as a competitor to standard rational choice models, I focus on how the theory of Problematic Situations goes beyond the assumptions of the latter.

Standard rational choice models assume stable intrinsic preferences (see Stigler & Becker 1977), that is to say that intrinsic preferences remain independent of the situation of

action. If I *intrinsically* prefer chocolate ice cream over strawberry ice cream, my preference will not change. At best, new information can lead the extrinsic preference to change, for example when I learn that the ice cream shop also offers chocolate ice cream and therefore buy that rather than strawberry ice cream.

In the case of negotiations, this means that agents have their intrinsic preferences prior to the negotiations and that these intrinsic preferences remain fixed during the process. This assumption underlies the concern that group agents such as states cannot serve as the problem-solvers needed for climate change and similar large-scale problems: if their intrinsic preferences lock the agents into a prisoner's dilemma, they seem to have no way out.

Pragmatist sociology, including the Negotiated Order approach, breaks with the assumption of stable intrinsic preferences. Instead, Problematic Situations prompt preference change, including the change of intrinsic preferences. In response to encountering a Problematic Situation, agents become more likely to undergo a change of their preferences. By the end of the second chapter we should have a grasp on what exactly occurs in such Problematic Situations. But to understand the full dynamic ensuing from Problematic Situations, we must also understand the importance pragmatist sociology assigns to interpretation and meaning.

For this purpose, the third chapter reconstructs the symbolic interactionist heritage of the Negotiated Order approach. The task proves difficult since "interactionism has never been concisely formulated" (Rock 1979: 7). Blumer offered what comes closest to a summary of the foundational principles (Blumer 1969: 2):

1. Human beings act towards objects, including other human beings, based on the meanings these objects have for them.
2. The meaning of objects arises out of social interaction between human beings.

3. The agents engage in interpretative processes towards the objects they encounter.

Since these principles remain vague, I am compelled to reconstruct them at great length. At the risk of committing a heresy, I use tools from rational choice theory for my reconstruction efforts. Rational choice accounts of conventional meaning and signalling capture much of what Blumer and other symbolic interactionists were after.

David Lewis' *Convention* (1969) offered one of the first rational choice models of meaning. Conventions, including linguistic conventions about meaning, arise out of particular cooperation problems. Brian Skyrms (1996, 2004, 2010) and others have further developed such approaches to meaning, drawing on evolutionary theory. These models of signalling can account for how meaning arises in interactions and how individuals interpret and act on it.

While I argue that rational choice models can cover much of what Blumer discussed, I focus on the original contribution symbolic interactionism makes over and above such models. What rational choice theorists describe as signalling not only affects choices, but also the intrinsic preferences underlying the choices. As I show in the third chapter, the theory of Problematic Situations allows signalling to influence preference change and to thereby lead to a deeper form of cooperation. This finding forms the second pillar of my reconstruction of the Negotiated Order approach.

The fourth chapter then presents my analysis of Negotiated Orders. Drawing on the theory of action reconstructed in the previous chapters, I provide the necessary and jointly sufficient conditions for exhibiting a Negotiated Order. Negotiated Orders are a special form of social order, shaped by negotiation in response to Problematic Situations. According to the Negotiated Order approach, bureaucratic rules cannot explain all features of a social order on their own. We need to take negotiations and informal agreements into account.

Because the negotiations respond to Problematic Situations, there is an increased chance of preference change while they take place. The kind of signalling about preference change discussed in chapter three plays a role during these negotiations. Engaging in trade-offs is not the only way to cross a motivational chasm in negotiations; a change of motivations serves as a bridge as well.

Chapter five combines the theory of group agency with the Negotiated Order approach. Strauss and other practitioners of the Negotiated Order approach have asserted the existence of group agents. Relying on their word, I do not myself offer a defence of group agency. The Negotiated Order approach, however, does not provide a theory of group agency. Being a sociological school rather than one of philosophy of mind, it says little about the theory of mind for groups. I fill this gap by arguing that Negotiated Order sociologists should accept a coarse-grained functionalism about minds. My argument proceeds by rejecting the most influential competitor: the interpretivist theory of group agency defended by List and Pettit (2011), and Tollefsen (2015).

In addition, I argue that the Negotiated Order approach puts limitations on what can be the functional realisers of propositional attitudes in groups. The on-the-spot agreements created by negotiations have to be part of the realiser, not just bureaucratic rules, formal procedures, and official declarations. This proposed limitation stands in contrast to the picture suggested by List and Pettit's work.

The sixth chapter returns to the original motivation for developing a theory of negotiating group agents. I deliver on my promises and show that in the case of climate change negotiations, my account of group agents offers a different perspective than standard rational choice models. For this purpose, I consider models of climate change at length. As I discuss in chapter six, group agents do not face a simple prisoner's dilemma during

climate negotiations, yet the situation they face might resemble one (see Gardiner 2004, 2011).

While I stay clear of defending any ethical or political positions on climate change, my account points out potential paths for influencing climate change negotiations to achieve cooperation. Assuming that group agents face a situation resembling a prisoner's dilemma during climate change negotiations, one way to leave this situation behind is if the group agents undergo a preference change in response to Problematic Situations. Since debates on climate ethics have been largely based on standard rational choice models, my diverging account should interest researchers in this field.

In addition to the six main chapters, the present thesis includes an appendix, which provides a more extensive interpretation of Dewey's relation to decision theory. This appendix reconstructs Dewey's theory of practical reasoning as a contribution to decision theory and helps to formalise motivational change. I will refer to the formal tools outlined in the appendix at multiple points in the thesis to provide a deeper understanding, but they are not a pre-requisite for understanding the general line of argument.

On Reconstructing the Negotiated Order Approach

Reconstructing the Negotiated Order approach for philosophical purposes is an endeavour that is far from trivial. I therefore want to point out a few complications regarding my attempt before jumping into it.

Strauss and other researchers did not establish the Negotiated Order approach by writing a manifesto providing clear principles guiding all future applications.[7] Rather, Strauss and

[7] The closest that Strauss ever came to setting down explicit principles might be his list of 19 assumptions in *Continual Permutations of Action* (Strauss 2014: 23-45). But these assumptions

his team plunged head first into empirical research, drawing from various sources and following sometimes conflicting intuitions. Even in his later book *Negotiations,* Strauss prefers to discuss research findings rather than abstract principles. Accordingly, my thesis considers the sources from which Strauss draws in his empirical work, particularly Dewey's pragmatist theory of action and Blumer's symbolic interactionism.

Arising out of a long tradition of pragmatist sociology, the Negotiated Order approach has become a multifaceted approach with a precarious unity. One can reconstruct this approach from various angles and achieve quite disparate results. The textual basis remains a far from univocal. My own reconstruction unifies the Negotiated Order approach by contrasting it with standard rational choice theory, that is, decision and game theory, assuming stable intrinsic preferences. This take on the Negotiated Order approach lends itself for discussing models of negotiating group agents since recent philosophical discussions of group agency have made use of standard rational choice theory but not its lesser-known rival.

Strauss himself characterises his differences with game theory as follows:

"One of its [game theory's] virtues is that it does focus attention *directly* on negotiations and on such related modes of activity as coercing, persuading, and manipulating; but it just as surely preconceives the nature of the social orders that game theorists study and interpret, as well as the possibilities for action within the interpreted orders." (Strauss, 1988: 72)

Both standard game theory and the Negotiated Order approach shed light on negotiations and related modes of activity, but Strauss believes that some of the assumptions underlying game theory limit the description of such activities excessively.

focus on action, while leaving out other aspects of the Negotiated Order approach, and not all of them are shared universally within the approach.

In the following chapters, I discuss how the Negotiated Order approach relaxes these assumptions, particularly the assumption that intrinsic preferences remain stable.

To make the contrast easier, my reconstruction of the Negotiated Order approach moves it as close as possible to rational choice theory. I use the notion of preference for my reconstruction and argue that much of the symbolic interactionist heritage can be captured using rational choice models of signalling. Many pragmatist sociologists might consider this move a heresy, but I provide textual evidence for my reconstruction and, in the end, show that the Negotiated Order approach has more to offer than standard decision and game theory. In effect, by moving the Negotiated Order approach closer than usual to rational choice theory, I only grant more to my imagined opponent.

While my reconstruction has a clear basis in the work of Strauss and other pragmatist sociologists, it remains selective. I want to note one omission: identity. Strauss' earlier book, *Mirrors and Masks*, written before he went on to found the Negotiated Order approach, discussed how individuals assume identities such as student or professor and how these identities shape interactions and vice versa.

Although I occasionally draw on *Mirrors and Masks* to support my interpretation of Strauss, I do not discuss the role of identity in the Negotiated Order approach. While taking identity into account might reveal further differences between the Negotiated Order approach and standard rational choice theory, for my purposes I can exclude its role. The contribution of the Negotiated Order approach becomes apparent without wading into the difficult territory of identity.

Every concise reconstruction of the negotiated approach will be selective. Other reconstructions might focus on identity and not include the notion of the Problematic Situation, although it has a clear influence on Strauss and other pragmatist sociologists.

My selection justifies itself by including the right elements to allow an account of negotiating group agents, which goes beyond standard rational choice models.

Chapter Two: The Problematic Situation

The first leg of my reconstruction of the Negotiated Order approach is a theory of motivational change. The founder of the Negotiated Order approach, Anselm Strauss, stated that many actions either have no clear goal or, if there is such a goal, "nevertheless over time something happens to this imagined goal" (Strauss 2014: 35). This chapter provides a discussion of the special something that happens. Strauss claims that means-ends schemes are often incapable of covering the whole dynamic of action, because "[b]oth ends and means may be reformulated in transit because unexpected results occur. Commitment, even to a major way of life or destiny, is subject to revision in process" (Strauss 1997: 38, see also Strauss 2014: 33).[8]

I reconstruct the Negotiated Order theory of motivational change as a theory of preference change. My reconstruction remains completely descriptive, that is, I avoid any normative evaluations of the motivational change processes described by Dewey and the pragmatist sociologists and instead only render their empirical claims acceptable for philosophical debate. This descriptive approach raises the question of the empirical adequacy of my approach. I defer here completely to the authority of the pragmatist sociologists. They have engaged in qualitative research (e.g. Strauss et al. 1963, see also Strauss 2014) and found the processes that I reconstruct to be empirically plausible.

Strauss, however, was not the first to talk of a revision in process initiated by unexpected events. The idea belongs to the pragmatist heritage of the Negotiated Order approach. Problematic Situations[9] play an important role in Dewey's pragmatism: they matter for

[8] I use the term "commitment" in a specific way later on, which is explained in the appendix and is to be distinguished from Strauss' use.
[9] As mentioned before, I capitalise the term to indicate its technical use, which is discussed later on.

his theory of action, his epistemology, and his political theory. Given this, it might be surprising that only in a few passages does Dewey provide an explicit account of Problematic Situations.

Generally, we can distinguish between epistemically Problematic Situations and motivationally Problematic Situations. In epistemically Problematic Situations, agents have an increased likelihood of changing their epistemic states such as beliefs, while in motivationally Problematic Situations, agents are more likely to undergo a change of their motivational states such as preferences. While pragmatists insist that these two types of Problematic Situations often coincide so that an epistemically Problematic Situation is also a motivationally Problematic Situation, at least conceptually we can distinguish between them.

My chapter proceeds as follows in presenting the theory of Problematic Situations. First I introduce them by drawing on Dewey's discussion in his *Logic: The Theory of Inquiry*, which treats *epistemically* Problematic Situations. Although I am interested in motivationally Problematic Situations, it helps to start with epistemically Problematic Situations since Dewey has provided more material on these.

After presenting Problematic Situations in this general way, I specify the three necessary and jointly sufficient conditions of motivationally Problematic Situations. Each condition characterises one stage in the encounter of a Problematic Situation. Each stage receives treatment in a separate section.

The first stage is the indeterminate situation, which is characterised by disruption and confusion and leads agents to identify a problem. The identification of the problem is the central feature of the second stage. Such identification then leads the agent to become open to change. Agents opening up to motivational change in response to a Problematic

Situation enter an exploratory phase during which they try out various courses of action and might change their preferences according to the experiences they undergo.

Having introduced these three stages, and in the course of doing so with the Problematic Situation, I contrast this theory of motivational change with rational choice theory. While the endorsement of intrinsic preference change distinguishes the account of motivationally Problematic Situations from standard rational choice theory, some non-standard decision theory models allow for such change as well. Cohen and Axelrod's model of adaptive utility bears a striking resemblance to the pragmatist account of Problematic Situations. A closer look, however, reveals important differences. In particular, the pragmatist account emphasises the exploratory nature of preference change.

The chapter ends by comparing of my account with a model by Cohen and Axelrod and an initial application of the proposed preference change theory to a social situation. As it turns out, the theory of Problematic Situations has important implications for game theory, which I develop in later chapters.

The Epistemically Problematic Situation

To provide a basic introduction to Problematic Situations, I present Dewey's account of epistemically Problematic Situations. An epistemically Problematic Situation leads the agent to open up to a change of her epistemic states, such as beliefs. An epistemically Problematic Situation, however, is not only characterised by this kind of change but also by the way the change arises.

According to Dewey, a Problematic Situation arises out of an indeterminate situation. But what is a situation in the first place? And what might characterise a situation as either

indeterminate or problematic? Drawing on the work of Barwise and Perry (1981: 669), we can say that a situation is characterised by a spatio-temporal location and a type, where the type represents the agent standing in relation to various other objects. If I kick a ball, the situation is of the ball-kicking type, since I stand to the object of a ball in the relation of kicking it.[10]

An indeterminate situation is "uncertain, unsettled, disturbed" (LW 12: 109),[11] because something disrupts the agent's activity. Dewey describes such a situation as one pervaded by "a unique doubtfulness" (LW 12: 109). Rather than in a kicking relation, the agent stands in relation to the objects in that she is confused about these objects by virtue of a disruption of her activity. Accordingly, the indeterminate situation is characterised, first by a disruption of the agent's activity, and second by the agent's ensuing confusion about the objects with which she engages.

In their discussion of symbolic interactionism, Hans Joas and Wolfgang Knöbl use the example of a computer that stops working to illustrate such situations (Joas & Knöbl 2009: 197). In the following, I use an extended version of this example: I am writing my thesis. Unexpectedly, my text processing program misbehaves. It just won't format the citations the way I want them formatted. So far, the computer has served me as nothing more than a glorified typewriter, but now it disrupts my activity.

In the first moment, I am merely confused since I expected the program to format the citations the way I intended and do not understand why it does not. My writing activity

[10] In many of my examples I will assume that the relation is of actively engaging in a certain way with the objects e.g. kicking the ball. But this is not a necessary condition.

[11] Here and in the following, abbreviations to quote John Dewey's *The Collected Works*, the numbers indicating the volume:
MW: Dewey, J. (1976). *The Middle Works, 1899–1924*. J. A. Boydston (ed.), Carbondale: Southern Illinois University Press.
LW: Dewey, J. (1981). *The Later Works, 1925–1953*. J. A. Boydston (ed.), Carbondale: Southern Illinois University Press.

has been disrupted and I encounter an indeterminate situation. The citations simply won't take the form I intended for them. The confusion of the indeterminate situation leads the agent to look for and identify a problem, thereby turning it into a Problematic Situation (cf. LW 12: 111).

My thesis writing situation presents a clear enough and specific problem: I cannot meet my goal of formatting the citations in such-and-such a way with the program. Having identified the problem, the situation stops being merely confusing and turns problematic: *something must be done about it*. In the epistemic case, the agent subjects the situation and its components to inquiry.[12] I try to learn more about how I can format my citations with the text processing program and read about the various functions of my program.

In addressing the epistemically Problematic Situation, the agent opens up to changing her epistemic states, in particular her beliefs. My beliefs about computers and text processing programs are up for grabs. I actively look for clues[13] about how citation formatting works and might go so far as to acquire a computer handbook, having opened up to changing my beliefs and to exploring various courses of action. For example, I try another writing program and different formatting solutions to figure out what is going on.

In the end, I might have to give up on my belief that I can format my citations this way, while acquiring new beliefs about the general topic. According to Dewey's description, I have tried out different courses of action following various hypotheses, first wildly clicking various tabs and items on my computer screen, then reading up on formatting

[12] Dewey's description of the Problematic Situation here is very close to Peirce's conception of doubt and inquiry: "The irritation of doubt causes a struggle to attain a state of belief. I shall term this struggle inquiry, though it must be admitted that this is sometimes not a very apt designation" (Peirce 1992: 114).

[13] In this context Dewey writes of the suggestions that the situations provide. For the role of suggestions, see Dewey LW 11: 114 and MW 6: 239.

citations and various text processing programs until I could make sense of the situation by holding appropriate beliefs.[14]

Based on this introduction to epistemically Problematic Situations, I propose the following necessary and jointly sufficient conditions for Dewey's technical notion of a "Problematic Situation":

A situation is a Problematic Situation for an agent if

(i) it arises out of a situation of unexpected disturbance of the agent's activity,

(ii) the agent identifies a problem, and

(iii) in response, the agent opens up to a change of mental states.

Later on, I discuss all three conditions at length for the special case of motivationally Problematic Situations. Here I provide a basic sketch using my example of epistemically Problematic Situations. However, the first two conditions do not differ between the epistemic and the motivational case. The difference lies merely in which mental states become open to change.

These three conditions for a Problematic Situation describe the three stages that characterise a Problematic Situation. The first condition describes the stage of the indeterminate situation, which comes before the Problematic Situation and gives rise to it. In the computer example, the indeterminate situation is limited to the first moment in which I am just confused that the program fails to format the citations in the way I had intended and expected. My thesis writing activity grinds to a halt.

The second condition describes the identification of the problem, which initiates the Problematic Situation. In the computer example, I identify the problem that my expected

[14] Compare this with Dewey's description of the process in his essay "How We Think" MW 6: 240-241.

course of formatting the citations proves infeasible, because the program acts contrary to my expectations. Prompted by this identification of the problem, I start an enquiry into text processing programs, which leads to the next phase outlined in the third condition.

We only have a proper Problematic Situation if in response to the identification of the problem the agent is opened up to a change of mental states. Such opening up characterises the stage following the Problematic Situation, which I call the exploratory phase. In my example, I open up to changing my beliefs about text processing programs and formatting citations. This opening up does not have to be under agential control, that is, the agent does not need to volitionally change her mental states. All that is needed is that the agent has an increased probability of a change in propositional attitudes, in this case the change of beliefs. To indicate that agential control isn't necessary, I often use the passive formulation.

By that point I have, of course, already changed my beliefs somewhat: I had to give up on my belief that I could simply format my citations in such-and-such a way without running into difficulties. But these belief alterations form only the preamble to the changes of mental states mentioned in condition (iii), because they occur before the agent is opened up by the situation. Only the further belief changes result from the inquiry prompted by identifying the problem. They arise because I read up on text processing and try out various things.

A Problematic Situation in the sense presented here and found in pragmatist sociology needs to have all three features: the unexpected disturbance, the identification of a problem, and the opening up which characterises the exploratory phase during which agents are more likely to change their mental states. I use "Problematic Situation" as a technical term derived from Dewey's usage. The technical term overlaps partially with

colloquial use of “Problematic Situation”, but my definition does not cover all standard uses.

Not every situation which we deem problematic arises out of an unexpected disturbance. A situation can be morally problematic without an unexpected disturbance preceding it. In a straightforward sense, an immoral action makes a situation morally problematic. The conditions for Problematic Situations, such as the demand for a disturbance, have nothing to do with this situation being problematic in a colloquial sense. However, Dewey clearly had something else in mind and the pragmatist sociologists picked up on his theory.

The three conditions hold for Problematic Situations in general, and one can differentiate between epistemically and motivationally Problematic Situations based on which mental states agents reconsider in the exploratory phase. For epistemically Problematic Situations, the exploratory phase concerns epistemic states such as beliefs. For motivationally Problematic Situations, the exploratory phase concerns motivational states such as preferences. A situation might be a Problematic Situation of both types, since agents often reconsider epistemic as well as motivational states at the same time.

However, in my reconstruction of the Negotiated Order approach I focus on motivationally Problematic Situations, because the motivational case diverges from standard rational choice theory.

The Motivationally Problematic Situation

Dewey’s theory of the Problematic Situation influenced how pragmatist sociologists think about motivational change. Strauss stays close to Dewey when he offers the following description:

"But the future is uncertain, is to some extent judged, labeled and known after it happens. This means that human action necessarily must be rather tentative and exploratory. Unless a path of action has been well traversed, its terminal point is largely indeterminate. Both ends and means may be reformulated in transit because unexpected results occur. Commitment, even to a major way of life or destiny, is subject to revision in process-at least until the final commitment of self-sacrifice." (Strauss 1997: 38)

The emphasis on the unexpected and the reformulation of ends and means in transit derives from Dewey's theory of the Problematic Situation. As suggested in Strauss' quote, motivationally Problematic Situations share the main features of epistemically Problematic Situations. The same three stages characterise the theory of the Problematic Situation.

First the agent finds herself confused in an indeterminate situation because an unexpected event disrupts her activity. Strauss' formulation of "unexpected results" refers to this state.

Second, the agent identifies a problem thereby initiating the Problematic Situation. This identification might be fairly vague at the beginning. In the computer example given above, I might at first identify the problem only as the fact that the text processing program fails to format the citations the way I intend. In the course of my engagement, my description of the problem becomes more fine-grained.

Third, identifying a problem opens the agent up to changing her mental states during an exploratory phase. In a motivationally Problematic Situation, the agent opens up to changing her motivational states. Agents may reformulate ends and means in transit, as Strauss observes. Often, we do not stop our activities altogether and then act with ready-made new ends after having come up with a new course of action. Rather we explore different courses of action by trying them out.

While the quote shows that Anselm Strauss, the father of the Negotiated Order approach, endorsed the theory of motivationally Problematic Situations, many gaps remain. What kind of disruption prompts a Problematic Situation? What does it mean to identify a problem, and what is the nature of the exploratory phase? In this section, I present the motivationally Problematic Situation in general before I turn to the three stages: disruption, identification of problem, exploratory phase.

I reconstruct the theory of motivationally Problematic Situations as a theory of preference change. This is a departure from Dewey and the pragmatist sociologists since they do not use the term “preferences”, mostly for historical reasons and perhaps to distinguish themselves from rational choice approaches. Nonetheless, two reasons speak for this move:

First, it simplifies the comparison with standard rational choice models, which rely on the concept of preferences. My reconstruction aims to show that the Negotiated Order approach has more to offer than the standard rational choice models. To accept the conceptual tools of standard rational choice theory is to grant the opposed view the choice of weapons.

Second, the actual research by Negotiated Order approach sociologists allows for an interpretation in terms of preferences. Although in their more abstract theoretical work, Dewey, Blumer, Strauss, and others shun the term “preference”, the actual descriptions of interactions invite an interpretation in terms of preferences. Take this section from Strauss describing the situation before a negotiation in a medical institution:

“The stage was set for a full round negotiation: The physician wanted responsibility and operational leadership to pass to another ‘responsible’ person; the psychologist wanted an 'intellectual' colleague; the nurse wanted someone to teach her psychodynamics, and

the social worker wanted an ally to help mitigate the physician's 'arbitrary power.'" (Strauss 1988: 118)

It is natural to interpret this passage as saying that, for example, the psychologist preferred a negotiation outcome that provided him with an intellectual colleague over one that does not. The concept of preferences readily captures what Strauss describes in folk psychological terms. Given such passages, it appears acceptable to reconstruct the Negotiated Order approach using the concept of preference.

I distinguish intrinsic preferences from extrinsic preferences. The motivational force of an extrinsic preference depends on other preferences.[15] For example, if I prefer to drink from the cup of water on my table because of my preference for satisfying my thirst over remaining thirsty, then my preference to drink from this cup is derivative of the preference for satisfying my thirst.

In mixed cases, my preference for some state of affairs over another is partially derived and partially intrinsic. I might prefer to drink a cup of tea because I prefer to quench my thirst, but I might also want to do it because I like to drink tea. For the sake of simplicity, I am sticking with the polar cases of either fully intrinsic or fully extrinsic preferences, but extending my discussion should pose no difficulty. Having defended and clarified my use of preferences, I present an example of a Problematic Situation.

Matilda visits an alpine village every summer. During her current visit, she finds a path up a hill she had failed to notice before. She wants to walk this new path.[16] In fact, she has an *intrinsic* preference to take this path over any other path. She sets off down the path,

[15] I distinguish intrinsic from extrinsic preferences by a difference in motivational force, but the literature often points to a difference in justification (e.g. Binmore 2009: 5-6). An intrinsic preference would then be one for which no further justification, or at least no further justification in terms of preferences, can be given. I assume that the motivational and the justificatory aspect of preferences go together.

[16] A similar scenario can be found in Anderson 2014.

but then suddenly finds herself confronted with an unexpected obstacle. A tree has fallen across the path and it is too big for her to climb over it. Thick thorn bushes on both sides of the path render even a small deviation from it difficult. The tree blocks the path and Matilda's activity. She has to stop.

Just as in the previously discussed computer scenario, Matilda experiences confusion in an indeterminate situation because an obstacle disrupts her activity. The complete confusion lasts for only a moment since Matilda quickly identifies the problem: a tree is lying across the path. The disruption and following identification of the problem opens Matilda up to a change of her motivational states. In this example, one motivational state presents itself as especially relevant: Matilda has an intrinsic preference to walk this new path. She preferred to take the newly discovered path over the others. This intrinsic preference now has an increased probability of changing. Matilda has opened up to preference change.

The occurrence of a Problematic Situation does not guarantee that intrinsic preferences change. Matilda might keep her preference for walking this particular path even if she finds it blocked. She might fetch a chainsaw or she might just wait until the fire brigade removes the tree without ever giving up her preference for walking this path over other paths. The Problematic Situation only renders an intrinsic preference change more likely than if no Problematic Situation had occurred (*ceteris paribus*). The theory of Problematic Situations is strictly probabilistic.

Matilda might have maintained another, purely extrinsic preference to walk the path. She might have taken this route because it seemed a good way to satisfy her preference to reach the top of the hill. The intrinsic preference would have been to reach the top of the hill, rather than to take a particular path. In this case, the Problematic Situation might move Matilda to lose her extrinsic preference to walk this path and look for another way

to follow her intrinsic preference. She might just choose another path that leads to the top of the hill. In this scenario, there would be no intrinsic preference change.

However, while the availability of an extrinsic preference increases the likelihood that only such an extrinsic preference changes, the probability that the intrinsic goal of reaching the top of the hill will be abandoned increases due to the Problematic Situation as well. Problematic Situations do not only affect means. They can affect all ends, as Strauss' quote with which I started the present section suggests.

According to Dewey and the pragmatist sociologists, all goals are within the reach of Problematic Situations, which gives action an open-ended character (Strauss 1997: 38). Whether Matilda had the intrinsic preference to walk up the hill or to take this particular path, encountering a Problematic Situation increases the chances of a revision of extrinsic and intrinsic preferences.

Matilda's case fulfils the three conditions of Problematic Situations:

(i) There is a situation of unexpected disturbance when Matilda encounters the tree.

(ii) Matilda identifies a problem: A tree blocks the path she wanted to walk.

(iii) Matilda opens up to a change of extrinsic and intrinsic preferences.

I elaborate each of these three conditions in the following three sections.

Note here that only the third condition makes the difference between epistemically and motivationally Problematic Situations. In the epistemic case, the agent opens up to a change in epistemic states, in the motivational case she opens up to a motivational change. Again, the opening up does not have to be under agential control. Matilda does not have to intentionally reconsider her set of preferences.

Many Problematic Situations turn out to be epistemically and motivationally problematic, and pragmatists emphasise this confluence of epistemic and motivational changes. Particularly for cases in which only extrinsic preferences change, the difference vanishes. If Matilda wanted to walk the path only because she preferred to reach the hilltop and this seemed the easiest path, acquiring the knowledge that a tree is blocking the path suffices for Matilda to lose the extrinsic preference.

The claim that extrinsic preferences change comes as no surprise. The notion of an extrinsic preference implies that they depend on beliefs about how the agent can realise the underlying intrinsic preferences. A change of intrinsic preferences is more interesting and I therefore focus on such cases in the rest of my thesis.

The theory of Problematic Situations postulates that extrinsic and intrinsic preferences can change. However, the fact that an agent encounters a Problematic Situation does not guarantee a motivational change. The encounter merely raises the probability of the preference change during the exploratory phase prompted by the identification of the problem. The motivational change takes place during an exploratory phase in which the agent tries out various courses of action. But before I discuss this third phase, I will turn to the disruption and identification of a problem that precede it.

Unexpected Disruption

An unexpected disruption stops activity and causes confusion. The disruption and the confusion characterise the indeterminate situation preceding the Problematic Situation. I have suggested as much already, but two major questions remained open: What gets disrupted, and what counts as an unexpected disruption?

When Matilda goes for a walk on a path and finds this path unexpectedly blocked so that she must stop, the event counts as a disruption. An obstacle unexpectedly disrupts an activity. Matilda literally stops walking when she comes upon the tree. She engages in an activity before discovering the tree, and although she had intended and expected otherwise, she ceases her engagement in the activity upon her discovery.

The example of Matilda concerns the bodily behaviour of walking. Dewey and the pragmatist sociologists focus on examples of bodily activity. However, despite the focus on outward behaviour, a disruption can also affect a purely mental activity. A telephone call disrupts me while I multiply two large integers in my head. The call leads to disruption and confusion characterising an indeterminate situation even though the activity was purely mental.

Nothing in the pragmatist literature rules out such a case of merely mental activity. The focus on cases of bodily behaviour is a result of the fact that pragmatists, including pragmatist sociologists, are interested in how reflective mental activities arise in the first place (see, for example, MW 14: 54). But emphasising that reflective mental activity often arises out of disrupted pre-reflective activity does not entail that the reflective mental activity cannot get disrupted as well.

We can now specify what gets disrupted: an activity of the agent. The activity can be bodily or mental and the expectations can be tacit. What counts as an unexpected disruption of such an activity? While it does not matter whether the activity involves outward behaviour or mere thinking, the unexpected nature of the disruption does matter.

Strauss introduces the terminology of “trajectory projection” (Strauss 2014: 55) for capturing the expectations associated with activities. Agents internally project the course of activity. Occasionally they form explicit and exhaustive means-end schemes, but often

they maintain only rough expectations about how things might go, which remain outside conscious awareness. Matilda tacitly expects that she can simply follow the path up to the hill. It turns out to be more difficult. An indeterminate situation arises from a disruption that goes against the expectations with which the agent acted.

For Matilda, disruption results from the unexpected difficulty in following one's desires. Just facing a difficulty satisfying a preference does not suffice for a disruption. While I am keenly aware that my preference to visit Japan this summer over staying at home will remain unfulfilled because of my limited budget, this difficulty does not disrupt my activity. I never began planning to go to Japan this summer. Even if my financial constraints were unknown to me, as long I did not start planning or acting in some way based on the assumption that I will go to Japan, no activity is unexpectedly disrupted.

In contrast, if I had started planning my trip to Japan and tried to book my flight, then my activity would have been disrupted by the fact that my credit card did not cover the expenses. I might then have entered a Problematic Situation. The disruption does not guarantee that I encounter a Problematic Situation, but it characterises an indeterminate situation leading up to a Problematic Situation.

Facing a difficulty in satisfying a preference is not sufficient for a disruption, nor is it necessary. The disruption could result from acquiring new information violating expectations of the agent without rendering preference satisfaction more challenging. Assume Matilda wants to take the path expecting it to lead her up the hill. However, some ways up the path she comes across a fork and neither of the two paths takes her up the hill. Matilda's desire was to walk the path, and encountering a fork does not create a difficulty in fulfilling this desire. Going either way suffices to fulfil the preference for going on this new path over other paths. Nonetheless, having to select a direction disrupts the activity. Matilda stops walking. She might identify a problem and turn the situation into

one that is problematic, or she might just shrug her shoulders and move in one direction, but as long as encountering the fork unexpectedly interrupts the walking, it is a disruption.

In all the previous examples, the disruption of the activity has been external. The agent engages in an activity, such as walking, and external objects, a tree or a fork in the path, disrupt it. Even in the example of my journey to Japan, my credit card is what puts a limitation on the activity. However, internal disruptions also occur.

Assume that you solve some difficult mathematical problems in your head and suddenly you forget the last value. Without the value, you stop calculating. You could go back and calculate or give up, but in either case your mental activity is disrupted by purely internal causes, not because a telephone call diverted your attention. If you identify a problem that increases the likelihood of preference change, then you have encountered a Problematic Situation.

Disruptions come in a variety of forms. For an indeterminate situation, the disruption must violate some expectation, interrupt some already ongoing activity and thereby cause confusion. As mentioned, the activity can be physical or mental and the expectations can be tacit. What matters is that the activity stops at least momentarily because of an event that goes against the agent's expectations. If the situation fulfils these conditions, then the agent encounters an indeterminate situation. It turns problematic when the agent identifies a problem that opens her up to preference change.

Identifying a Problem

The indeterminate situation ends and the Problematic Situation starts when Matilda identifies a problem and this identification opens her up to change (cf. LW 12: 111). What exactly is Matilda doing when she identifies a problem in the situation of the blocked path? Does Matilda make a judgement that could be mistaken and if so, what makes the

judgement true or false? My account needs to clarify whether Matilda engages in a cognitive judgement and if so, what the truth-making facts for the judgement are.

Dewey writes in his discussion of Problematic Situations that "the situation is taken, adjudged, to be problematic" (LW 12: 111). This quote and in particular the term "adjudged" supports a cognitive analysis of identifying a problem. To identify a problem is to make a judgement about the situation, and so the identification could be wrong.

However, Dewey also wants this first judgement of identifying a problem to remain thin. Identifying the initial problem stands at the beginning of a longer inquiry "[t]o find out *what* the problem and problems are which a Problematic Situation presents to be inquired into, is to be well along in inquiry" (LW 12: 112). Only the initial identification is a necessary condition for a Problematic Situation. The further specification of the problem happens during the inquiry prompted by the Problematic Situation. In other words, the agent turns the situation into one that is problematic by making a judgement, but we need a judgement with rather thin truth-makers, that is truth-makers which do not require the agent to settle on too much.

To develop the thin notion of identifying a problem, I introduce a thicker one for contrast: the notion of identifying a well-defined problem. I show that such a notion demands too much and then return to the thinner notion. To identify a well-defined problem, we must specify the goal, the initial state, admissible operations, further constraints, and the outcome state. As Nozick helpfully clarified, the required goal serves as "an evaluative criterion for judging outcomes and states" (Nozick 1993: 164) of the problem. They are the standard by which to evaluate any putative solution to the problem. It is the specification of the goal which creates trouble.

Take Martin, who stands at the foot of a hill and has the goal of reaching the top as soon as possible. Martin picks a quick path up the hill and encounters an obstacle just as

Matilda did. Assume that Martin immediately identifies a well-defined problem. A tree blocks this path completely and there is no way around it. Martin specifies walking the available paths as admissible operations.[17] Among the further constraints are Martin's bodily capacities and the conditions of the paths. Given his lack of fitness, Martin cannot run up the hill all the way. The outcome state should be Martin standing on top of the hill. Most importantly, the goal is to reach the top as soon as possible. Unless Martin commits to a specification of the admissible operations, the constraints, the outcome state, and the goal, he has not identified a well-defined problem.

By identifying a well-defined problem, however, Martin closed down the option of changing his intrinsic preference for reaching the hilltop. Endorsing a goal is necessary for having a well-defined problem. To specify a well-defined problem for oneself entails a commitment to a goal and therefore to certain preferences. In the example, Martin sticks to his intrinsic preference for reaching the top of the hill, and only changes the extrinsic preferences concerning the route to take.

Identifying a well-defined problem already narrows down the potential motivational change. A problem is well-defined only relative to a fixed goal. Accordingly, endorsing a well-defined problem rules out giving up on the goal that specifies the well-defined problem and adopting new goals that outweigh it. To completely open up to changing his goal of reaching the top of the hill would entail that Martin gives up his well-defined problem. On Dewey's account, in contrast, the identification of the problem is supposed to open the agent up to preference change rather than to close that possibility down. To open up to changing his goal, Martin would have to give up his well-defined problem, not endorse it.

[17] I take here the notion of an admissible operation to be the same as that of an option worth considering.

Identifying a problem in Dewey's broader sense is a crucial step to opening the agent up for a motivational change, including a change of intrinsic preferences. Identifying a well-defined problem always puts some goal beyond the reach of the situation. Finding the path blocked prompts Matilda to reconsider her intrinsic preference for walking this path. Does she really prefer to take this way over any other way? She opens up to changing the one intrinsic preference we stipulated her to have. No goal pertaining to the situation is beyond the reach of the Problematic Situation.

The notion of a well-defined problem subordinates the changes to a fixed goal, but Problematic Situations can affect all goals after the identification of the problem. In other words, the necessary conditions for identifying a problem cannot require that one considers a goal as being fixed. At least in principle, a Problematic Situation could turn all preferences upside down.

This is not to say that in all Problematic Situations, all goals open up to reconsideration. When I encounter a Problematic Situation trying to format the citations in my thesis, I open up to changing my preference, but I will not reconsider my goal of writing a thesis. This Problematic Situation does not unsettle me enough to have such an effect.

In sum, it is not a necessary condition of Problematic Situations that no goal is considered fixed, but the necessary conditions must allow for situations which reach all goals. By identifying a well-defined problem, Martin already narrows down the motivational change in a way that puts goals beyond the reach of the situation. The notion of "identifying a problem" in condition (ii) for Problematic Situations has to remain thinner. Nozick, to his credit, is well aware that we encounter problems in which our goals "may not be predetermined; a person may select his goals or alter the ones he has been given" (Nozick 1993: 165 Footnote).

Earlier I asked what the truth-making facts for identifying a problem could be, and the discussion of the well-identified problem provides us with the following answer: a judgement that identifies a problem can be true even without a fixed goal relative to which something is a problem. Matilda does not have to correctly specify a fixed goal to make the judgement of identifying a problem true.

Instead we should see the identification as the agent making the judgement that she just faced an indeterminate situation. Only confusion caused by an unexpected disruption of activity characterises the indeterminate situation, not judgement. The identification of a problem is the judgement of having encountered an unexpected disruption of one's activity. The truth-making fact of the judgement is not that the situation poses a well-defined problem, but rather that the agent's activity cannot proceed as expected, is interrupted in an unforeseen way, and therefore confusion ensues.

Matilda identifies the problem that she cannot proceed with her activity as she preferred and expected. She wanted to walk a path and since a tree lies across the path, and she can neither easily climb over it nor walk around it, the activity is disrupted. While I claimed above that to identify a problem, the agent does not have to be committed to a goal, Matilda's motivation to walk the path comes into play here after all. This preference, however, becomes opened up for reconsideration. There is no *fixed* goal, because the difficulties in following the preference and judging the existence of such a problem have unsettled the agent's goal.

One might suggest that Dewey's theory assumes a fixed goal relative to which the situation is taken to be problematic, even if this time the goal looms in the background. The goal would be more general, perhaps the goal to have preferences which allow uninterrupted activity, or the goal to grow as a person through experience (cf. Hook 1959).

Following this proposal, Matilda identifies a problem insofar as she judges that preferring this path over others fails to allow uninterrupted activity.

The pragmatist sociologist can respond to this suggestion in two ways. One option is to accept that there is such a goal relative to which the situation is problematic, but it is the goal of a second-order preference. The preference for uninterrupted activity might be the higher-order preference to develop certain first-order preferences. The other option is to deny that the agent needs such a second-order preference to identify the problem, and instead argue that the agent engages in a less instrumental form of reasoning. We should endorse the second option and deny the need for a second-order preference, which is not to rule them out for all cases.

The pragmatist literature does not provide the basis for attributing such a second-order preference. Dewey emphasises the primacy of activity (MW 14: 84-87), but he does not mean to say that uninterrupted activity is the highest goal from which we derive others. Rather, Dewey claims that activity is a default state for agents to which we usually return after disruption. After a disruption, human agents return to activity because this has been hard-wired into them. They do not need a second-order preference to motivate them, since activity serves as a fallback option. Nothing I found in the pragmatist literature sufficiently supports postulating a second-order preference instead.

Since I am reconstructing the pragmatist Negotiated Order approach, Dewey's word has a pro tanto authority. Other things being equal, I should stick to his presentation. However, if it turned out that only by postulating a second-order preference can one make sense of the Negotiated Order approach, then a rational reconstruction should postulate it after all. I must show how we can do without a second-order preference. Concretely, why should we call it a problem if it is not a problem relative to a goal the agent is still committed to? This question concerns the justification of our terminology.

I suggest that to identify a problem, the agent does not need an instrumental inference relying on a second-order preference. That the agent registers that her activity has not gone its projected course but rather has been disrupted justifies calling the judgement "a problem identification". The agent already had a motivation that set the standard by which to evaluate whether there is a problem.

Matilda walks up the path only to find her way blocked. She identifies a problem since she wanted to walk this path to the top of the hill and a tree stopped her from doing so. Her activity has been disrupted and this prompts a change to her preferences, including the preference that originally put her on this path. But she does not have to engage in an instrumental inference of the form:

a) I have the goal to engage in undisrupted activity.

b) My activity is disrupted because of my preferences.

c) Therefore, the situation is problematic regarding my preferences.

Rather, Matilda identifies the situation as problematic because the activity does not work out as expected. The disruption suffices for the problem without a higher-order preference against disruptions.

Consider the parallel case of finding preference fulfilment good. You do not need a preference for having your preferences fulfilled to find it good that your preference has been fulfilled. You just need the preference, which has been fulfilled. Matilda can find it good to walk the path in virtue of having an intrinsic preference to walk this path rather than another. Likewise, you do not need a preference for having your preferences fulfilled to find it bad that your preference has not been fulfilled. You just need an unfulfilled preference.

Finally, for you to find a problem in how your preference led you into a situation in which things are not working out, you do not need a preference that your preferences should

work out. You just need the preferences which led you to this situation. Importantly, the agent does not have to remain committed to the preferences which led to the situation. Having followed one's preferences, to find one's activity unexpectedly disrupted suffices to open the agent up to preference change on the pragmatist model.

Earlier I asked why we should call it a problem if it is not a problem relative to a goal that the agent is still committed to. We can call it a problem because the agent engaged in some activity that was disrupted against the agent's expectations. Something went wrong, therefore the agent faces a Problematic Situation even though there remains no goal to which the agent is fully committed and relative to which the situation is problematic.

These considerations do not rule out that some agents in fact have such second-order preferences that come in during or after Problematic Situations. However, the analysis of Problematic Situations, in particular condition (ii), does not require such preferences. The agents can identify a well-defined problem and a problem relative to a second-order preference, but neither identification is necessary for a Problematic Situation.

A situation is a motivationally Problematic Situation only if an identification of a problem in the required sense leads the agent to open up to a motivational change; that is, only if it increases the likelihood of preference change.[18] How exactly the preferences change depends on the developments in the exploratory phase following the problem identification. Hence, the next section discusses the exploratory phase.

[18] This is the empirical claim, which I reconstruct from the pragmatist basis of the Negotiated Order approach. I do not defend its empirical adequacy but rather rely on the theory's endorsement by pragmatist sociologists of motivational change to support it.

Exploratory Phase

The unexpected disruption of activity and the identification of a problem are necessary conditions for a Problematic Situation. Strictly speaking, opening up to preference change is a necessary condition, but not the following exploratory phase. In principle, an intervention could stop the agent after opening up and before exploring. A Problematic Situation opens Matilda up for preference change, but if a bolt of lightning hit her, she would never go through an exploratory phase.

Nonetheless, the exploratory phase deserves our attention. Although it is not a necessary condition for a Problematic Situation, it distinguishes the pragmatist theory of preference change. As we will see later, it is lacking in other models of preference change and the next chapter develops the important role of the exploratory phase.

In the exploratory phase, the agent tries out different courses of action without yet having fully committed to the motivation underlying this course of action. Agents can simulate the exploration in their imagination, but they also engage in exploratory behaviour. The exploratory phase is a period of time in which the agent acts with a low commitment on various potential motivations. Having opened up to preference change, the agent acts with tentative motivations and the feedback she receives influences whether she actually settles on this motivation.

Having tentative motivations, the agent follows the course of action she would engage in if she endorsed certain preferences, but without yet having settled on them completely. I propose that to have a tentative preference is to have a preference with a limited commitment.[19] Discussing Problematic Situations, Dewey writes that "the decision reached [is to be regarded] as hypothetical and tentative until the anticipated or

[19] I formalise this idea and further specify the notion of commitment in the appendix.

supposed consequences which led to its adoption have been squared with the actual consequences" (MW 12: 173).

In the exploratory phase, the agents maintain preferences in a qualified state, that is, with limited commitment. Not all preferences might be held in this state, but at least those the agent associates with the identified problem will be.[20] These preferences are tentative. The agent has a high chance of giving them up again if the consequences suggest it, a topic which I address at length in the next chapter.

Matilda does not have a firmly established preference when she chooses a new path after encountering the Problematic Situation. She tentatively tries adopting new preferences, giving up preferences, and changing their strength without yet fully committing to these changes. Eventually, however, the agent settles on a new motivational profile. The commitment to the preferences rises and the exploratory phase ends for the agent. She effectively loses some preferences and endorses new ones.

One might ask what leads the agent to settle on one motivational profile rather than another at the end of the exploratory phase. In the epistemic case, the agent draws inferences to establish a new profile of beliefs. Does the agent in the motivationally Problematic Situation become more likely to engage in some form of practical inference?

For an alteration of extrinsic preferences, the answer is yes. Assume that Matilda wants to get to the top of the hill, finds the way blocked, and enters a Problematic Situation. In response, she gives up her extrinsic preference to go up this path to the hill rather than using another path. She keeps her intrinsic desire to get to the hilltop, however. This intrinsic preference to get to the top of the hill together with the belief that the next best

[20] The exact scope of which preferences are qualified is a major open question of the Negotiated Order approach. For my purposes, I leave it at the criterion of an association with the problem (see also the appendix).

path to walk is on the south side of the hill, lead her to the practical inference that she should go to the south side. But in this case the inference rests on stable intrinsic preferences and the instances in which these preferences change interest us more.

Can an inference also establish a new intrinsic preference? At the end of the exploratory phase, does the agent infer the new motivational profile including all changes to intrinsic preferences? The question raises a thorny issue, which will return: Which forms of practical inference are valid, and whether agents can ever validly infer intrinsic preferences, are disputed matters.[21] Following the pragmatist sociologists, I maintain that Problematic Situations can lead to new intrinsic preferences. *If* I proposed that the agent established her new motivational profile by way of valid practical inferences, then I would be committed to the claim that agents validly infer intrinsic preferences.

Perhaps to the dissatisfaction of some philosophers, I sidestep the issue. The license for my evasion follows from my aim: reconstructing the Negotiated Order approach and applying it to negotiating group agents. This aim is descriptive. The Negotiated Order approach as developed by Strauss belongs to sociology, not a theory about which forms of practical inference are valid.[22]

While I promise a theory of preference change, my theory remains descriptive and operates at a relatively general level. The process through which agents change their motivational profile might be a form of inference and this form of inference might be valid or not. In fact, the agent might undergo a mix of inferential and non-inferential changes, and the inferences might be a mix of valid and invalid forms of inferences. I

[21]For one theory that postulates something close to an inference to intrinsic desire, see Millgram 1997; see also the edited collection Millgram 2001 for related issues in the debate on practical reasoning.

[22]Dewey might have thought that he discussed a form of *valid* practical inference, but I am mainly interested in the sociological contributions.

merely consider how to describe the motivational changes independently of these concerns.

When an agent changes her beliefs in response to an epistemically Problematic Situation, the agent might also engage in a fallacious theoretical inference. The agent draws new inferences, and rationality requires her to draw only valid inferences, but a descriptive theory can and should allow that agents make mistakes. Even if an agent engages in fallacious reasoning, we would still want to describe the situation that led her to reconsider her beliefs. Descriptive models of epistemically and motivationally Problematic Situations can leave the question of validity to others.

Nonetheless, I reconstruct an informative model: It makes the prediction that in certain circumstances agents are more likely to change their preferences, including their intrinsic preferences which rational choice theorists (e.g. Stigler & Becker 1977) consider as unchanging. In the next chapter, I offer even further predictions. These will clarify how the tentative preferences of the exploratory phase settle in and become preferences with a full commitment.

I have developed all three conditions for Problematic Situations: the unexpected disruption leading to the identification of a problem, which opens the agent up to test new motivations in a qualified manner during the exploratory phase. Before I move beyond individual cases in my reconstruction of the Negotiated Order approach, I compare the theory of Problematic Situations with another model of preference change.

Comparison with Cohen and Axelrod's Adaptive Utility

By introducing intrinsic preference change I deviate from the rational choice models usually employed in the social sciences, and especially economics. Stigler and Becker

(1977) argued that economic models can do without the change of intrinsic preferences and many followed their suggestion, including the rational choice models of climate change negotiations between states.

Some non-standard rational choice models, however, allow for preference change and one might wonder whether the Negotiated Order approach offers something different from them. From the available models[23] I select a particularly relevant one: The model of adaptive utility developed by Cohen and Axelrod in their 1984 paper "Coping with Complexity: The Adaptive Value of Changing Utility" resembles my Deweyan account while offering a helpful contrast. It will establish that one can construct mathematical models similar to the pragmatist theory of Problematic Situation and at the same time helps to see what makes the latter unique.

Cohen and Axelrod intend to establish that preference change can be adaptive. They want to offer a model "where such a process [of changing preferences] can improve performance" (Cohen & Axelrod 1984: 30). Along a relevant dimension of assessment, the outcome has to improve because the intrinsic preferences change in response to the situation. The agent acts *better*, in a sense to be further specified, because she changes her intrinsic motivations. Cohen and Axelrod defend their view with an example:

A factory manager has to divide a fixed number of labour hours between the production and the maintenance of the machines. Being a factory manager, she wants to maximise output. She aims to achieve this goal using an expected output function. This is the output function she believes to be valid; wrongly as we will see. The expected output function

[23] Theories and accounts of preference change include but are not limited to Dietrich & List 2011, 2013, Hansson 1995, Grüne-Yanoff & Hansson 2009, Hansson & Grüne-Yanoff 2012, Dekel et al. 2007.

specifies the expected output ($\hat{y}$) using the hours of labour devoted to production (x) and a parameter b, with the expected value $\hat{b}$.

Because there are multiple rounds of production, the values need time indexes. The expected output, the expected value of b, and the productive labour vary over each round. The manager estimates the expected output $\hat{y}_t$ using the estimate $\hat{b}_{t-1}$ from the previous round of production and the actual number of labour hours devoted to production in the current round (x_t). Putting all this together, Cohen and Axelrod provide the following expected output function:

$$\hat{y}_t = -x_t^2 + \hat{b}_{t-1} x_t$$

This function describes what the manager *believes* to be the relation between the hours of production labour and the parameter b. She uses this equation to guide her actions. Cohen and Axelrod also stipulate an equation for how the manager estimates b:

$$\hat{b}_t = \frac{y_t}{x_t} + x_t$$

For each round, the manager calculates her estimate of b using the output of this round (y_t) and the labour invested into production during this round (x_t). The manager then uses this updated estimate in the next round of production.

However, in Cohen and Axelrod's example, the manager's estimated output function does not match the actual output function. The manager overlooked another parameter, namely the output lost to pilferage. They include this parameter as c in the output function, and since it tracks the pilferage, it takes a value below zero ($c < 0$).[24] The correct output function is:[25]

[24] They also assume for sake of simplicity that this factor remains stable.
[25] Cohen and Axelrod also considered other production functions and came to similar results, see Cohen & Axelrod 1984: 38.

$$y_t = -x_t^2 + bx_t + c$$

So far Cohen and Axelrod specified how a factory worked and how the manager expected it to work without including preference change. The next two steps introduce preference change.

First, Cohen and Axelrod stipulate that the factory manager receives intrinsic utility for labour devoted to production. We calculate this intrinsic utility by multiplying the hours of labour *x* with a value *w*, which we can interpret as the intrinsic utility of one hour of production labour.

Since the manager also receives utility for the output (*y*), we end up with the following utility function:

$$U_t = y_t + w_t x_t = -x_t^2 + (b + w_t)x_t + c$$

The function is nothing more than the sum of the production output with the intrinsic utility of the productive labour. Because the manager does not know the amount of output before letting the factory do its work, we also have a function for the expected utility:

$$\hat{\mathrm{U}}_\mathrm{t} = \hat{\mathrm{y}}_t + w_t x_t$$

The expected utility is the expected output plus the intrinsic utility of the productive labour.

Second, Cohen and Axelrod stipulate that surprise governs the value *w.* In other words, surprises change the amount of intrinsic utility that the manager derives from one hour of labour devoted to production. Cohen and Axelrod propose measuring the surprise the manager experiences by subtracting the expected utility for one round of production from the actual utility of that round:

$$D_t = U_t - \hat{\mathrm{U}}_t$$

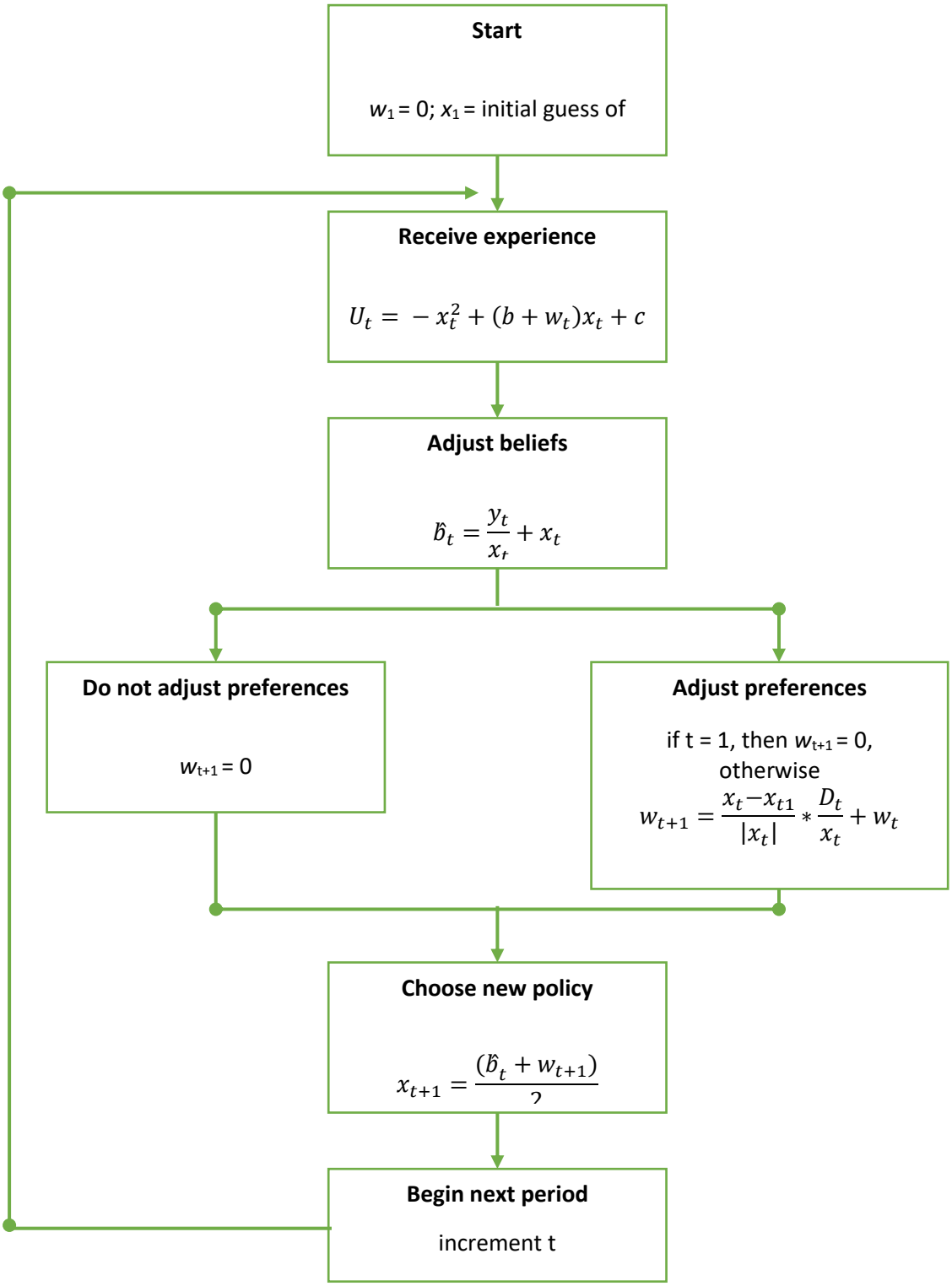
The surprise governs the intrinsic utility associated with productive labour. All we now need for a model of preference change is an equation specifying how *w* changes over time. Cohen and Axelrod suggest the following equation:

$$w_{t+1} = \frac{\frac{x_t - x_{t-1}}{|x_t|} * D_t}{x_t} + w_t$$

This equation makes surprise the main driver of preference change and puts a drag on the change so as to rule out runaway preferences (see Cohen and Axelrod 1984: 35).

To answer the question as to whether this equation describes actual preference change in humans, we would need empirical evidence. However, Cohen and Axelrod only want to show that under certain conditions, preference change improves the performance of the agent. Whether human agents in fact show such preference change remains a secondary consideration.

The following graphic (taken with minor adaptations from Cohen and Axelrod's paper) provides a comparison of how the models differ with and without preference change:


Start
$w_1 = 0$; x_1 = initial guess of
Receive experience
$U_t = -x_t^2 + (b + w_t)x_t + c$
Adjust beliefs
$\hat{b}_t = \frac{y_t}{x_t} + x_t$
Do not adjust preferences
$w_{t+1} = 0$
Adjust preferences
if t = 1, then $w_{t+1} = 0$, otherwise
$w_{t+1} = \frac{x_t - x_{t1}}{|x_t|} * \frac{D_t}{x_t} + w_t$
Choose new policy
$x_{t+1} = \frac{(\hat{b}_t + w_{t+1})}{2}$
Begin next period
increment t

Cohen and Axelrod calculate that for a wide range of values, the agent with preference change makes choices superior to those of the agent without. But first we have to clarify the benchmark used to determine that the agent's choices are indeed superior. Because the preferences change, we cannot compare the quality of the choices using the amount of utility. If we did that, an agent with a large *w* would appear superior to the agent who always sticks to *w* = 0, even if the factory produced very little or nothing at all. The factory manager would hardly excel if she just followed her idiosyncratic preferences and let the business slide.

In effect, Cohen and Axelrod only compare the output of factories.[26] Since *w* always takes a positive value in the equation, if the preference changing agent creates higher output, she also receives higher utility. The manager who not only cares about output but also about hours of labour devoted to production, achieves more output than the manager who cares only about output.

This surprising outcome results for many values, because the managers cannot exactly predict output. The parameter, *c*, escapes the manager and allows the adaptive preference model to have an edge. That managers lack perfect insight into all variables influencing the production process is probably the least worrying assumption Cohen and Axelrod make for their model. We can question the psychological plausibility of the equation describing the change of *w*, but the limited knowledge of production processes should raise little controversy. This plausible assumption makes the difference between the performances with and without preference change.

While the result of the model provides a rationale for preference change, my interest concerns the striking resemblance it bears to Dewey's theory of Problematic Situations.

[26] Cohen and Axelrod (1984: 36) introduce more complex considerations, but the output is what the difference boils down to.

In both theories, unexpected outcomes lead to a change of preferences. Preference change happens in response to a situation arising from limited knowledge. We can adapt the example of Matilda to the assumptions of Cohen and Axelrod's model with only minor modifications.

Assume that Matilda has an intrinsic preference to walk a certain path. She receives the utility a_t if her preference is fulfilled. In addition, she has an intrinsic preference to reach the top of the hill, leading her to receive utility b_t if her preference is satisfied. The intrinsic preference to walk the path exceeds the preference to reach the top of the hill ($a_t > b_t$), but Matilda believes that the path should lead to the top of the hill, so that the expected utility of the walk is a_t+b_t.

She finds to her surprise, however, that the path goes around the hill rather than to the top. The discovery of the different route leads Matilda to stop and identify a problem. The path does not lead where she thought it would. This new information surprises Matilda. We can even calculate the surprise as the difference between the expected and the actual utility as suggested by Cohen and Axelrod: it is exactly b_t since that was the intrinsic preference of reaching the hilltop.

The surprise and the identification of a problem might lead to a change of her intrinsic utilities, so that a_t becomes a_{t+1} and b_t becomes b_{t+1}. If the utility change is big enough, so that $b_{t+1} > a_{t+1}$, then Matilda would take another path up the hill. The preference for a path up the hill now guides her action. Encountering the Problematic Situation leads to a change of intrinsic preferences.

I have specified the example so that it fits with Cohen and Axelrod's model as well as the pragmatist account of Problematic Situations. Just as in Cohen and Axelrod's example, I have given a preference which changes in response to surprise. At the same time, the example fits Dewey's theory since Matilda encounters a Problematic Situation and

undergoes a change of intrinsic preferences in response. Both theories postulate intrinsic preferences adapting to the unexpected. There remains, however, significant and illuminating differences between the approaches.

Differences

While Cohen and Axelrod describe a mathematical toy model to show that agents with preference change can outperform agents with stable intrinsic preference, Dewey and the pragmatist sociologists provide a complex but informal theory of how agents undergo actual motivational changes. As a result of these varying interests and approaches, and in spite of all convergence, numerous differences emerge. Some of these differences hardly matter in the present context. For example, Cohen and Axelrod talk about utility, while I have couched Dewey's theory of Problematic Situations in terms of preferences. Little rests on this for present purposes. I want to discuss two more consequential differences that shed light on the theory of Problematic Situations.

First, the theory of Problematic Situations and Cohen and Axelrod's model differ insofar as the pragmatist theory lets a disruption and the identification of a problem prompt the opening up to preference change, while Cohen and Axelrod let *surprise* change preferences.

While the two mechanisms of disruption/identifying a problem and surprise resemble each other and can give rise to the same results, as seen in the adapted Matilda example, key differences remain. According to Dewey, only a disruption of activity in which the agent identifies a problem increases the probability of preference change. If a surprising disruption occurs but the agent identifies no problem, then she does not change her preferences, even though she has been surprised.

For example, Matilda might find a £50 note on her walk. She stops for a moment to pick it up. The discovery disrupts her activity and surprises her, but she does not see any problem. She puts the money in her pocket and walks on without opening up to preference change. Finding the money, however, leads her to experience more utility than expected from taking her walk. The money is a welcome surprise and Cohen and Axelrod's model suggests that such a surprise should shift the preferences. Accordingly, the predictions of the two approaches diverge.

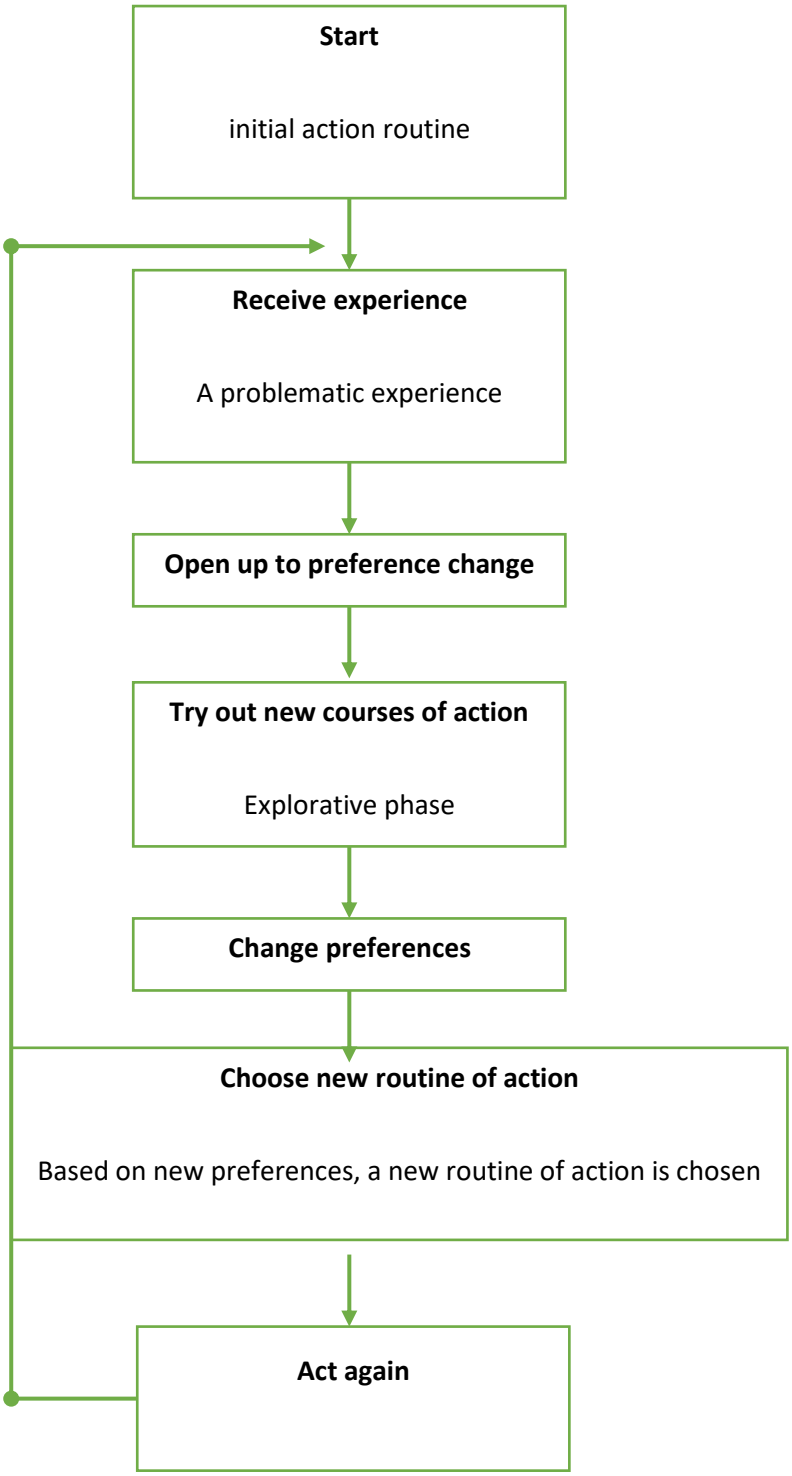
While surprises and Problematic Situations often take place as one package, we should not confuse them. An unexpected disruption is part of the analysis of Problematic Situations, and if we were to count every unexpected event as a surprise, then a surprise belongs to the necessary conditions for a Problematic Situation. Since Cohen and Axelrod operationalise surprise as the difference between expected and actual value of a variable, they seem to endorse a concept of surprise as violation of expectations.

But surprise is insufficient for a Problematic Situation, and therefore Dewey offers a more restrictive account of preference change. All situations of unexpected disruption are situations of surprise, but not all of them lead to Problematic Situations. Since economists have disputed the very existence of intrinsic preference change, restricting the scope of the category is a sensible move. In any case, I am reconstructing the Negotiated Order approach following Dewey. Psychological experiments could test the empirical adequacy of the pragmatist account versus Cohen and Axelrod's model, but this is not my project. I take the pragmatist sociological approach as given and reconstruct it as charitably as possible.

Second, Cohen and Axelrod portray preference change as instantaneous, while Dewey and the pragmatist sociologists emphasise the exploratory nature of such processes. According to the theory of Problematic Situations, agents explore different courses of

action before a new motivational profile is firmly established. Matilda looks both ways and takes steps in multiple directions before settling on a new motivational profile.

Cohen and Axelrod's model involves no probability and instead assumes the immediate development of new preferences. They also lack anything analogous to the exploratory phase. The factory manager never holds any tentative preferences exploring various courses of actions, but changes her preferences immediately. We can show the difference between the models by adapting the flowchart used above to now describe the pragmatist model (without equations):



Cohen and Axelrod's model does not include the box "try out new courses of action". Instead of opening up to preference change, agents immediately change their preferences. I introduced tentative preferences characterised by a limited commitment to capture the exploratory nature of action emphasised by Dewey and Strauss. By

contrast, Cohen and Axelrod's manager either has a preference or not, without any variation in commitment.

These significant differences result from the problems of modelling preference change. Including probabilities and an exploratory phase in their model would have complicated it without contributing to their aim. As mentioned, Cohen and Axelrod intend to establish that preference change can be adaptive and to do that they can assume a simplified picture of such change.

I am interested, however, in how agents negotiate with one another, and for this purpose the exploratory phase proves important. The exploratory phase allows agents to explore each other's motivational landscape while having a low commitment to their own preferences. The next chapter discusses the impact of the exploratory phase on the interaction between agents. Before that, I take a first step towards introducing preference change into models of interdependent action of multiple agents.

From Individuals to Game Theory

The theory of preference change in response to Problematic Situations has implications for game theory. The models of interdependence resulting from taking the theory of Problematic Situations into account differ from standard game theoretic models. To convey this difference, I will consider the well-known prisoner's dilemma.

In a standard prisoner's dilemma, the result of the game affects two agents who cannot enter binding agreements. Each of them prefers the outcome of mutual cooperation over the result of mutual defection. However, each of them most prefers an outcome in which the other cooperates but they themselves defect. The worst outcome results from the other defecting, while they themselves cooperate. Given this pay-off structure, defection

becomes the dominant strategy. Defection is the move which maximises preference satisfaction regardless of what the other player does.

Prisoner's Dilemma		Player B	
		Cooperate	Defect
Player A	Cooperate	2, 2	0, 3
	Defect	3, 0	1, 1

Assuming agents act according to these fixed preferences, mutual defection inevitably results. By introducing the capacity to change preference in response to Problematic Situations, however, the result stops being inevitable. The theory of Problematic Situations allows the transformation of the prisoner's dilemma, as the following example illustrates:

David and Simon had hoped to organise a conference together, but the prospects appear dismal. Currently they are each on separate flights to their annual joint meeting with their funding body, so they cannot communicate. To get the conference funding from their funding body, at least one of them must read a batch of papers before the meeting. Only reading the papers will give them the necessary information to present the topic of the conference well. If neither of them reads the papers, they will not receive the funding and the conference cannot take place. If both read the papers, they present the topic even better and secure slightly more funding.

However, Simon and David could also work on their thesis or on their next publication. Since they are indifferent as to the choice between working on their thesis or on their next publication and both options imply defection from reading the papers, for now we can group the two options into one: defection. Later, however, it matters that there are

two ways of defecting. Having sketched their options, we need to specify how they rank them.

David and Simon prefer mutual reading over mutual defection. But both prefer to defect while the other does the reading over both doing the reading. Finally, they prefer mutual defection over doing the reading alone. As a result, their pay-offs are that of a prisoner's dilemma:

PD		David	
		Read	Defect
Simon	Read	2 (Simon), 2 (David)	0, 3
	Defect	3, 0	1, 1

We should expect both Simon and David to get off their flights without having read the papers. The conference appears doomed.

However, something unexpected occurs. Simon starts to work on his thesis after take-off, but he then discovers that he left an important file back home on his USB flash drive. Confused, he looks for the flash drive in various pockets to no avail. He doesn't have the file he needs for continuing to work on his thesis. The inaccessibility of the file disrupts his activity and Simon identifies a problem: He cannot go on writing on his thesis without the file.

Given his prior preferences, it is clear what Simon should do now: start writing his next publication.[27] He was indifferent as to the choice between working on his thesis and

[27] The example assumes, of course, that the USB flash drive is not needed for the publication.

writing the next publication, and stuck in a prisoner's dilemma with regard to reading the papers. Given that working on his thesis is now out of question, his prior preferences suggest turning to the publication.

But Simon finds himself in a Problematic Situation and opens up to preference change.[28] Maybe working on his next publication could turn out to be not so important for him and he might instead invest his time in reading the papers. In an exploratory phase, Simon starts looking into the papers. He explores this different course of action and finds that it suits him after all. His motivations changed and the encounter with the Problematic Situation results in different pay-offs:

		David	
		Read	Defect
Simon	Read	4 (Simon), 2 (David)	2, 3
	Defect	3, 0	1, 1

As one can see, reading rather than defection now dominates Simon's options. Accordingly, he reads the papers, Simon and David receive their funding, and the conference takes place. The preference change transforms the game and a different outcome emerges.

This result is almost trivial. Hardly anyone doubts that if agents preferred different options, the outcome would differ. However, the contribution of the Negotiated Order approach is to provide guidelines for when preferences change. Confusion and the

[28] As I noted before, the Negotiated Order approach does not offer a detailed specification as to which preferences open up to change. I presume here that the relevant preferences for the situation become tentative.

identification of a problem, which have both received attention in the present chapter, precede the motivational changes.

Furthermore, this application of the theory of Problematic Situations provides only a hint of what is yet to come. In the outlined scenario, Simon undergoes a preference change while on a plane – one that is in isolation from other relevant players. While Simon's motivational change affects the interaction with David, and so paves the way for organising a conference together, Simon and David do not interact during the exploratory phase before the new motivational profile settles in. In the next chapter, I argue that interaction during the exploratory phase makes a difference for preference change.

Chapter Three: Symbolic Interactionist Roots

Dewey's account of Problematic Situations is not the only influence on Strauss' Negotiated Order approach. The approach grew out of symbolic interactionism and accordingly shares the latter's basic assumptions. In the present chapter I contribute to the reconstruction of the Negotiated Order approach by discussing these assumptions and extending the account of preference change to social contexts. While symbolic interactionism comes in various flavours, we can focus on the Chicago version spearheaded by Herbert Blumer, since it had the most influence upon Strauss, who learned sociology in Chicago.

I begin my reconstruction in this chapter by presenting three principles Blumer put forward in his "The Methodological Position of Symbolic Interactionism", included in the 1969 collection *Symbolic Interactionism: Perspective and Method*. These principles emphasise agency, meaning, and interpretation. Agents act based on the meaning that objects, including other human beings, have for them, and engage in complex and flexible processes of interpretation which reshape the meaning of these objects. I act towards the raised hand of a student in my seminar based on the meaning it has for me, in this case the meaning of asking permission to speak. The interpretation process can be quite complex, for example, if I know that a student has already contributed a lot and therefore refrain from picking her. Such processes can then reshape the meaning of raising a hand: the good student's hand develops a different meaning than those of others.

Blumer takes his three principles to distinguish symbolic interactionism from other approaches. But I ask whether rational choice accounts might meet all three principles. The theories of conventions and signalling developed by such authors as David Lewis (1969) and Brian Skyrms (2010) capture Blumer's discussion of agency, meaning, and interpretation while endorsing standard rational choice theory. Blumer's meaning

becomes the signalled information and introducing common knowledge allows David Lewis to give the information a conventional and flexible nature and emphasises interpretation. Even the importance of the interpretation process has its place in David Lewis' account of conventions. Apparently, core principles of symbolic interactionism dissolve into standard rational choice theory.

Symbolic interactionists, however, can give signalling a role which the assumptions of standard rational choice foreclose. Since symbolic interactionists follow Dewey, they can draw on the theory of Problematic Situations, which I reconstructed in the previous chapter. This theory describes a mechanism of motivational change, including the change of intrinsic preferences, while standard rational choice approaches, such as the models by Lewis and Skyrms, do not allow for such change.

In the penultimate section of the chapter, I argue that signalling, i.e. the meaning and interpretation of meaning, affects the process of preference change. Because agents enter an exploratory phase in which they try out different potential motivations, they are open to intervention by other agents. The signal of one agent to another can affect the way the other's motivations change. This allows agents to align not just their actions by way of signalling, but also their intrinsic preferences. The standard rational choice approach with its assumption of fixed intrinsic preferences has nothing analogous to offer.

The chapter ends by addressing complications resulting from extending signalling accounts to model preference change.

The Principles of Symbolic Interactionism

The project of reconstructing symbolic interactionism poses many difficulties. Symbolic interactionism never achieved complete theoretic unity. In his *The Making of Symbolic Interactionism,* Paul Rock notes that "[s]ince its very beginnings interactionism has never

been concisely formulated" (Rock 1971: 7). Because symbolic interactionists preferred empirical research over abstract theory, they have been wary of spelling out their principles. Instead, they founded their school upon an oral tradition originating in the Chicago School. As a result, symbolic interactionism has fuzzy boundaries, overlaps with other traditions of sociology, and contains conflicting positions.[29] To quote Rock again:

"Because its tenets have not been clearly laid down, because its contours and boundaries are imprecisely drawn, there cannot be any exact demarcation of the place of interactionism within the larger territory of sociology." (Rock 1971: 16)

No authoritative manifesto codifies the tenets of symbolic interactionism. One text, however, comes close. In his 1969 "The Methodological Position of Symbolic Interactionism", Herbert Blumer presents the three core principles of symbolic interactionism. Although these principles do not capture all the contributions of symbolic interactionism, they resemble the necessary and jointly sufficient conditions for an account of interaction to fall within the school. Blumer introduces the principles in the following passage, calling them "premises":

"Symbolic interactionism rests in the last analysis on three simple premises. The first premise is that human beings act toward things on the basis of the meanings that the things have for them. […] The second premise is that the meaning of such things is derived from, or arises out of, the social interaction that one has with one's fellows. The third premise is that these meanings are handled in, and modified through, an interpretative process used by the person in dealing with the things he encounters." (Blumer 1969: 2)

Meaning and its role in human action lie at the heart of the three principles. Humans act based on meaning, which arises out of interaction and which humans interpret. I act

[29] This did not get better in the decades following Rock's work: see Fine 1993. Snow 2001 even argued for broadening the tradition beyond Blumer's three principles.

towards the student who raises her hand in my class based on the meaning students and raised hands have for me.

Blumer's three principles are:

1. Human beings act towards objects based on the meanings these objections have for them.
2. The meaning of objects arises out of social interaction between human beings.
3. The agents engage in interpretative processes towards the meaning of the objects they encounter, which reshapes the meaning.

In the following three subsections, I go through each of these three principles and discuss the restrictions they provide for an account of interaction to fall within the boundaries of symbolic interactionism. In the next section, I will attempt to reconstruct symbolic interactionism using rational choice theory.

First Principle

The first principle states that human agents act towards objects based on the meaning these objects have for them. Objects should be broadly construed to include humans, events, and situations. Blumer, however, remains unclear as to the meaning of "meaning". That he did not have primarily linguistic meaning in mind is beyond dispute, but without a further interpretation of "meaning", the principle remains opaque.

Two interpretations suggest themselves: first, the meaning of the object could be the full mental representation of the object, including all associations with the object. The meaning of a book might then be that I know that is a book, that I remember buying it, and that I associate a certain pleasure with the experience of reading it.

Second, the meaning of the object could be the information conveyed by the object. In the case of a book, it could convey information on its pages. Given as a present, a book on writing style conveys the information that the other person thinks I could improve my writing. I discuss the notion of information further in the next section.

As we can see, one can distinguish a mentalist and an information theoretical analysis of Blumer's concept of meaning. I endorse the information theoretical analysis because it fits better with Blumer's discussion of Mead's theory of gestures.

According to Blumer's reading, a "gesture is any part or aspect of an ongoing action that signifies the larger act of which it is a part" (Blumer 1969: 9). Blumer illustrates this with the example of "a robber's command to his victim to put up his hands [which] is (a) an indication of what the victim is to do; (b) an indication of what the robber plans to do, that is, relieve the victim of his money; and (c) an indication of the joint act being formed, in this case a holdup" (ibid.).

All three indications singled out by Blumer describe the informational content of the robber's command. The utterance explicitly contains the information that the victim is supposed to put up his hands, but there is also other implicit information, such as that the robber plans to acquire the victim's money. Blumer's discussion of the second principle supports interpreting "meaning" as the information an object conveys. We can reformulate the first principle: human beings act towards objects based on the information these objects convey to them.

One might consider this principle rather trivial, but remembering the historical context challenges this impression: radical behaviourism still held sway in psychology when Blumer first put forward his principle. Radical behaviourists denied the relevance of information processing by the agent insofar as they denied that there were any mental entities with which the agent engaged internally.

The first principle distinguishes symbolic interactionism from radical behaviourism, but also from other approaches. Blumer attacks psychologists who "turn to such factors as stimuli, attitudes, conscious or unconscious motives, various kinds of psychological inputs, perception and cognition, and various features of personal organization to account for given forms or instances of human conduct" (Blumer 1969: 3). Of course, Blumer does not deny the importance of perception and cognition for action. He is aware that without perception, objects could hardly convey information and without cognition, little information processing could occur. Blumer objects to neglecting the meaning of objects over these other factors, without denying that one can study both.

Objects convey information and the agents absorb this information, process it, and act accordingly. Raising your hand in my seminar conveys the information that you intend to ask a question. Perception plays a role here, but does not suffice on its own. I know that you want to say something not only because I perceive your hand above your head, but because I have the necessary background knowledge. Cognition alone does not suffice either; it has to process the information received from the hand within the social context.

As Blumer is well aware, symbolic interactionism is not alone in giving meaning such a guiding role; a cognitive psychologist will also accept that agents receive information and that they act based on this information. The second principle differentiates symbolic interactionism further.

Second Principle

According to the second principle, meaning arises out of social interaction, rather than the object itself or the psychology of one individual. Blumer elaborates:

"The meaning of a thing for a person grows out of the ways in which other persons act toward the person with regard to the thing. Their actions operate to define the thing for

the person. Thus, symbolic interactionism sees meaning as social products, as creations that are formed in and through the defining activities of people as they interact." (Blumer 1969: 4-5)

The meaning of raising your hand in a seminar arose out of social interactions. At one point the convention was established through interactions, and students were taught how to use the gesture.[30]

One might worry, however, that not all information arises out of social interaction. Robinson Crusoe on his island notices that one flower blooms shortly before the rain period. The blooming of the flower now conveys to him the information of the coming rain period. The meaning of the flower has changed.[31] How then should we make sense of Blumer's claim that meaning arises out of interaction?

Since Blumer does not address this issue directly, I propose the following reconstruction: we should not understand the claim that meaning arises out of social interaction as a universally quantified statement claiming one source for all meaning, but rather as a generic statement for the cases of meaning, which are core for Blumer, in particular, shared meaning in social situations. Blumer is not interested in providing a perfectly general analysis of meaning. He is not an information theorist trying to analyse the nature of information in all domains. He instead emphasises the importance of information conveyed *in social interaction* for his sociological research. Social interaction is a key contributor to establishing shared meaning, but not a necessary condition.

As a sociological school, symbolic interactionism has a special interest in shared meaning. Shared meaning is information, which the object conveys to multiple persons who have

[30] Like information, conventions receive more attention in the next section.
[31] The same problem arises for the mentalist interpretation of what Blumer means by "meaning". Robinson Crusoe has a mental representation of the flower as announcing the coming rain.

common knowledge of this fact. For example, when a student raises her hand in a seminar, the information that she wants to speak is shared by everyone in the room who sees the hand, and we all have common knowledge of this situation and the hand's meaning. Symbolic interactionism addresses the question of how meanings "become taken-for-granted and routinized" (Snow 2001: 372), that is, how the information of signals becomes robustly established in social practices.

Many symbolic interactionist studies turn around the ways multiple individuals create shared meanings. These studies discuss how information transfer becomes shared and stable for multiple agents. The shared meaning typically arises out of interaction since interaction establishes the common knowledge needed for the information to be shared. Many Negotiated Order studies discuss the *negotiation* of shared meaning (e.g. Maines 1982, Bryant & Stensaker 2011) as one type of interaction establishing shared information.

The second principle demands that the typical information, which is the focus of sociologists, arises out of social interaction. This principle still does not suffice to distinguish symbolic interactionism from various other approaches. Blumer aims to get more specific.

Third Principle

Blumer's third principle emphasises interpretation. Not only do human agents act according to the meaning of objects, where the meaning arises out of interaction, but they also engage in complex interpretation processes. Taking the notion of interpretation in a broad sense, it would be trivial to say that an agent interprets the information in her environment. Every agent, even a cockroach scrambling into a dark corner, processes information. Interpretation, in Blumer's sense, entails more than information processing.

Blumer's third principle serves to emphasise the role of the agent regarding meaning: "The actor selects, checks, suspends, regroups, and transforms the meanings in the light of the situation in which he is placed and the direction of his action" (Blumer 1969: 5).[32] The agent engages flexibly with the information and the interpretative process affects the meaning of the objects.

While a cockroach exhibits some flexibility by fleeing into the dark, it does so to a lesser degree than Blumer demands.[33] Human agents show great flexibility in their actions. When the students raise their hands in my class, I do not just pick the first one I see, but instead consider the circumstances. Who spoke last? Who looks most confident? My reaction depends on various further considerations. I am flexible in how I interpret and respond to the conveyed information.

Flexible agency is not all; in addition, Blumer demands conscious awareness. As he writes, during interpretation, the agent "has to point out to himself the things that have meaning" (Blumer 1969: 5). Blumer believes that "[t]he capacity of the human being to make indications to himself gives a distinctive character to human action" (Blumer 1969: 15).[34]

While sub-conscious information processing occurs within human agents, the kind of interpretation Blumer emphasises requires conscious engagement with the information. Blumer and the symbolic interactionists in general focus on the meaning of objects, insofar as their research subjects' conscious awareness makes a causal contribution to the interpretation. This puts a further constraint on what Blumer means by "meaning". He does not just mean "conveyed information", but "consciously accessed and processed

[32] See also Snow 2001: 373-374 on the principle of human agency in symbolic interactionism.
[33] I'm not committed to Blumer being right on this difference between animal and human agents. However, symbolic interactionism traditionally draws such a contrast. For a critical discussion of this aspect, see Alger & Alger 1997.
[34] While Blumer thinks that only humans and no other animals engage in interpretation of the kind discussed here, this does not form the core of symbolic interactionism and therefore is not part of my reconstruction.

conveyed information". The cockroach presumably processes information only in the first sense, without conscious access.

The third principle demands that the conveyed information goes through conscious and flexible interpretation processes. Blumer makes a further claim, which belongs in the scope of the third principle: interpretation plays a role in the transformation of meaning in interaction. According to Blumer, "interpretation should not be regarded as a mere automatic application of established meanings but as a formative process in which meanings are used and revised as instruments for the guidance and formation of action" (Blumer 1969: 5). The interpretation processes reshape the information conveyed by the object in the future.

A student raises her hand signalling to me that she wants to speak. However, she contributes frequently to the seminar. I give her a slight nod of the head and a smile, acknowledging her effort and knowledge, while taking another student who has spoken less. My response is not automatic but flexible and reflective. This interaction also changes the information the same action has in the future. The next time the student raises her hand only to show that, in principle, I can pick her to speak if no other student comes forward. Both of us share knowing glances. She conveys different information with the gesture in virtue of the previous interaction. Our flexible interpretation practices, that we do not always interpret a raised hand as a wish to speak and respond automatically, lead to a change of informational content.

Such examples illustrate Blumer's take on conscious interpretations and their power to reshape the meaning of objects through interaction. They are everyday examples described from a folk-psychological perspective.

In summary, Blumer's three principles demand

1. that consciously accessible information guides actions,
2. that in the typical sociologically relevant cases the information arises out of social interaction, and
3. that flexible and conscious interpretative processes shape how the information feeds into action. The interpretations and the associated interactions change the informational content.

We can consider a theory which meets these principles to fall within the boundaries of symbolic interactionism. In the next section, I undertake the attempt to offer a rational choice reconstruction of symbolic interactionism.

The Rational Choice Account: Signalling and Common Knowledge

At the time Blumer formulated his three principles, game theory failed to satisfy his demands for a theory of interaction. Game theory models typically assumed all-knowing agents, who responded mechanically to given incentives. Interpretation processes played no role worth mentioning in guiding action. The agents might have seemed to be limited automata, incapable of processing information in the flexible manner envisioned by Blumer. However, in 1969, the same year Blumer's "The Methodological Position of Symbolic Interactionism" appeared in his collection *Symbolic Interactionism: Perspective and Method*, David Lewis published his seminal book *Convention*, which introduced conventions and signalling games to rational choice theory.

In recent years, games of signalling and information processing more generally have become well-established areas of research. In a series of books, Brian Skyrms (1996, 2004, 2010) has presented key insights from these developments. With these contributions, standard rational choice theory offers an account of interaction meeting Blumer's principles.

First Principle

According to Blumer's first necessary condition, meaning guides human actions. Objects in a situation have meaning, insofar as they convey information in situations. Brian Skyrms defines informational content as "how the signal affects probabilities" (Skyrms 2010: 34). That the student raises her hand in class raises the probability that she wants to say something from the teacher's point of view. Skyrms' information allows an analysis of Blumer's meaning within the framework of game theory.

Not all Skyrms-type information counts as Blumer's meaning, however. Skyrms (2010: 29-30) engages in evolutionary game theory and discusses cases in which bacteria transfer information to coordinate. They release certain molecules and the reception of a certain quantum of molecules increases the probability of enough bacteria being in a certain state. While of interest to Skyrms' evolutionary project, such examples of information transfer between bacteria are far removed from what Blumer had in mind. As discussed in the previous section, information must be consciously accessible to the receiving agent to count as meaning in Blumer's sense. The meaning of an object is not determined by all ways it affects probabilities but only by those ways which are consciously accessible to the receiver.

The occurrence of smoke increases the probability of a fire taking place and therefore conveys information in Skyrms' sense. The smoke has the meaning of fire only if the agent perceiving it consciously accesses the information of likely fire. My proposal is that Blumer's notion of meaning picks out a subset of Skyrms' information, namely the consciously accessed information.[35] While Blumer limits himself to consciously accessible

[35] Blumer's focus on our consciousness also suggests that we are dealing here with subjective probabilities. The student's raised hand raises my subjective probability of her wanting to speak.

information, game theory can very well accommodate this restriction. There is no principled conflict between the two approaches here.

Second Principle

In his *Convention*, Lewis intended to analyse the sense in which linguistic meaning is conventional. As Quine and others had pointed out (cf. Lewis 1969: 1-4), linguistic meaning cannot be conventional in the sense that people met to discuss and agree what each word should mean. People did not convene to establish a convention. Linguistic meaning must get off the ground without presupposing linguistic exchanges. In response to this attack on the conventionality of meaning, Lewis turned to game theory. He conceived the problem of establishing conventional meaning as a coordination problem.

A coordination problem is a game with multiple coordination equilibria, that is, a combination of strategies "in which no one would have been better off had any one agent alone acted otherwise" (Lewis 1969: 14). For example, the two of us have to meet at a certain time in one of our offices. It does not matter to either of us in which office we meet, but it matters that we end up in the same room, since neither of us has the time to run to the other office if we failed to meet at the same one. The pay-off matrix looks as follows:

CP	Office 1	Office 2
Office 1	1, 1	0, 0
Office 2	0, 0	1, 1

No strategy dominates this game, because no move promises to be the best irrespective of the action by the other player. Instead, the game confronts us with two coordination

equilibria, the cases where we meet in the same office. We just need to coordinate. In the usual cases of meeting in an office, we could verbally agree on one place; but since Lewis is interested in how linguistic meaning arises in the first place, I leave this option aside for the moment.

A Lewisian convention provides a way to achieve coordination without linguistic communication. With a convention to meet in my office rather than yours, the problem dissolves. The convention provides the solution to the coordination problem, but understanding the incentives of the coordination problem enables us to see how we establish a convention without explicit agreement, that is, without presupposing linguistic meaning. Our interests already align. We want to have a convention so we do not miss each other in the wrong office.

Lewis saw that in a repeated coordination problem, finding oneself following a behavioural pattern in which an equilibrium is realised provides a reason to stick to the pattern. If we have started to always meet in your office, then we both have reasons to go there for the meeting. I expect you to be in your office and you expect me to go there.

Based on this insight Lewis offered an analysis of conventions. Simplifying slightly, a regularity in the behaviour of agents in a recurrent game is a convention, if and only if it is common knowledge amongst the players that in the instances of the recurrent game,

1. almost everyone conforms to the regularity,
2. almost everyone expects almost everyone else to conform to the regularity,
3. almost everyone has approximately the same preferences regarding the possible combinations of actions,
4. almost everyone prefers to conform to the regularity, on condition that almost everyone else conforms to it,

5. almost everyone would prefer to conform to another regulation, on condition that almost everyone was to conform to it (cf. Lewis 1969: 78).

The two of us follow the behavioural regularity of meeting in your office and we expect each other to do so. Our preferences are approximately the same and in favour of keeping to the regularity as long as the other does.

The incentive structure of a coordination problem, behavioural patterns, and common knowledge of the structure and the patterns bear the brunt of Lewis' analysis of conventions. Lewis introduced the influential analysis of common knowledge as a principally infinite hierarchy of "*i* knows that *j* knows that … knows that *A*" (cf. Vanderschraaf & Sillari 2013). I know that you know that we generally meet in your office and you know that I know that, and so on.[36] Given this common knowledge, we both have a reason to show up at your office rather than mine.

The term "common knowledge" is a bit of a misnomer as Lewis admitted, since a structure of hierarchical iterated *beliefs* suffices for giving a reason to follow the conventional pattern (cf. Lewis 1978: 44, Footnote 13). If I believe that you will be in your office and you believe that I believe and so on, this gives me a reason to go there, although we only believe rather than know. However, since the term "common knowledge" has become the standard, I will stick with it. We should only keep in mind that the requirement for common knowledge remains thinner than the term suggests.

Often common knowledge arises out of a public announcement, and therefore linguistic meaning plays a role. However, as Lewis (1969: 56-58) discusses, common knowledge of the conditions can also arise by other means, such as past experience of conforming to

[36] Since Lewis wrote, further analyses of common knowledge have been proposed. I omit here the debates about whether Lewis' hierarchical analysis is the right one and whether there has to be a highest level of knowledge or belief. As long as game theorists can use one analysis of common knowledge my argument stands.

the regularity. I notice that we always meet at your office in the past and you notice it too. We infer that we are both aware of this regularity and that we both prefer it over another one as long as the other keeps to it. Since common knowledge only requires beliefs, we can easily establish common knowledge in this case.

With the analysis of convention in place, we can turn to convention-based signalling. Lewis proposes that signalling games are coordination games. Consider a traffic light at a road junction where our cars meet. If we manage things such that we all stop on red and drive on green, we survive. But we also survive if we all stop on green and drive on red. We face a coordination problem. A convention saves our lives.

Traffic lights convey their information to stop or drive only because of the behavioural patterns of drivers. While fire causes smoke under a variety of conditions independently of any agent, traffic lights cause others to drive or stop depending on our previous interaction. The convention gives the traffic lights meaning, in Blumer's sense. Only by virtue of the convention does the red light convey the information that I should stop, and the convention depends on previous interaction.[37] The previous interaction establishes the behavioural pattern and the common knowledge of it.

In contrast to the fire scenario, this conventional meaning arises out of interaction just as Blumer would have it. Our behavioural regularity forms a condition for the convention, which allows the traffic light to convey the information we access. If everyone drove at red and stopped at green, the convention and accordingly the informational content would change. Lewis' analysis of convention captures this dependence on interaction and the rational choice approach accordingly conforms to Blumer's second principle of symbolic interactionism. Key cases of shared meaning arise out of interaction.

[37] There might be some natural reasons, why we should use red. Perhaps it is easier to see. However, such details at best render the coordination game slightly impure.

Common knowledge helps to explain how information arises out of interaction. Only because I know that you know that red lights serve as signals to stop, do I drive on green without slowing down. It is the special sauce which allows rational choice models to increase the informational content of signals. This common knowledge-based rational choice proposal bears a striking resemblance to pragmatist sociology in the tradition of Mead. Take the following passage from Hans Joas summarising Mead's position:

"[H]umans anticipate the way partners in action would potentially behave in response, and create an inner representation of that response. This ability enables humans to gear their behaviour to what potentially would be that of partners. As the partner, assuming it to be a human being, also had the same ability, a completely new pattern in the history of evolution emerges for coordinating behaviour: coordination by means of a shared orientation towards patterns of mutual behavioural expectations." (Joas 1996: 187, see also Blumer 1969: 9-10)

Ignoring the dubious claim about the evolutionary uniqueness of humans, this quote could re-describe David Lewis' finding. The drivers coordinate their interaction on the road based on mutual expectations. A new coordination arises because two agents have a hierarchy of beliefs or knowledge. In this light, using common knowledge appears in line with later pragmatist sociology and allows us to account for various phenomena that symbolic interactionists have traditionally focussed on.[38]

Symbolic interactionists might worry that the rational choice reconstruction only covers the special case of coordination problems, while Blumer does not only want meaning to

[38] Joas' summary of Mead has a slightly more behaviourist tendency than the suggestion by the rational choice approach. The quote qualifies the expectations as behavioural, while Lewis introduced common knowledge in mentalist terms. However, my reconstruction of the Negotiated Order approach is more cognitivist than the original pragmatist sources of Dewey and Mead. This cognitivist reconstruction facilitates the use in sociology. Mead might write about behaviour, but Blumer focusses on meaning and conscious interpretation.

guide action in the case of traffic lights and the like, where the preferences of the agents align. Meaning also guides the actions of agents in conflict. If the rational choice models were limited to coordination games of aligning interest, then they would fall short of the demands made by symbolic interactionists.

Coordination problems play an important role in the rational choice discussion of signalling, especially when it comes to explaining linguistic conventions. They deserve this attention, because their multiple equilibria allow for an element of arbitrariness in conveying information. Stopping on green and driving on red serves us just as well as the other way round. But coordination problems are not the only game where common knowledge and signalling make a difference.

Take the classic example of burning bridges in game theory. A general leads her troops across a bridge into a battle and has the bridge burned behind them. The bridge would have allowed easy transportation and there is no equivalent substitute. If the troops win the battle, they would prefer a bridge behind them for transportation. If they lose the battle, it is better to have a good retreat opportunity. Under such conditions, burning the bridge might seem irrational. The general might know, however, that she is playing a game of chicken, also known as the Hawk-Dove game, against the enemy troops.

This game of chicken has the following four outcome states. If both sets of troops fight with full force without retreating, both will be utterly destroyed and no one will win. However, if one troop fights with full force and the other retreats early, the first one will be victorious and suffer mild damage, while the second will lose and suffer mild damage. The pay-offs can be represented in the following matrix:

Chicken	Fight	Retreat
Fight	-1, -1	2, 0
Retreat	0, 2	1, 1

Burning the bridge removes part of the infrastructure for the retreat and increases the cost for one player. We could represent it in the matrix as follows:

Chicken	Fight	Retreat
Fight	-1, -1	2, 0
Retreat	-.5, 2	1, 1

For the row-player, that is the player whose options are represented in separate rows, retreat becomes a worse option. Looking at the new matrix we might be puzzled as to why anyone would burn the bridge as it only worsens one's options. However, the matrix is not everything, because it does not show an associated shift of expectations. We must consider how burning the bridge functions as a signal.[39]

The impact of destroying the bridge being common knowledge, setting it on fire conveys important information: the general burning the bridge intends to fight to the end. If she considers retreating, burning the bridge is not a good move. That she sets it on fire indicates that she will commit all the force of her troops even if it results in her own demise. The signal conveys information, not about the pay-offs, but about the probability that a player will choose a certain act.

The common knowledge of the intention to commit all forces changes what is rational for the other general. Given the pay-offs as specified before in the matrix, retreat becomes the only rational option. Letting the other win and allowing oneself to survive is better than mutual destruction.

Common knowledge changes the information conveyed. Only because the troops march towards the battle, and only because the structure of the situation is common knowledge, does burning the bridge signal a resolve to fight to the end. As we can see, meaning arises

[39] In effect, we have here a case of costly signalling, a type of signalling I discuss later.

out of interactions and common knowledge in situations other than just coordination problems.

As I argued, this rational choice approach meets the first two principles Blumer proposed for symbolic interactionism: Agents act according to the meaning of objects and this meaning arises out of interaction, at least for core cases. What about the third principle emphasising interpretation?

Third Principle

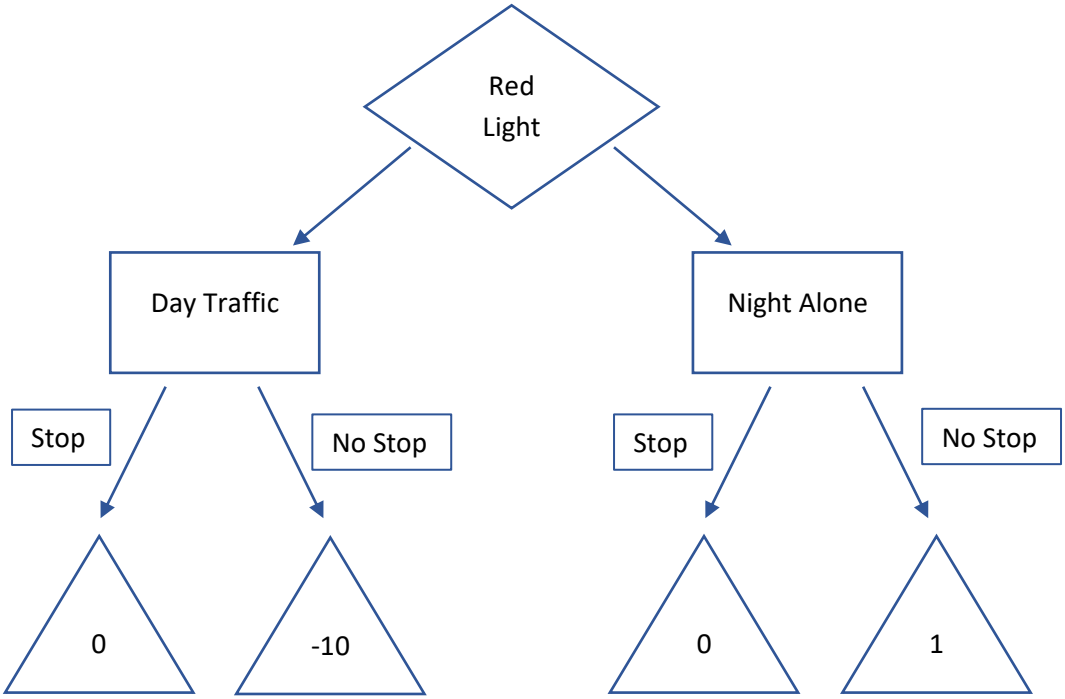
As mentioned before, Blumer has something special in mind when he emphasises interpretation. The generals engage in a limited form of interpretation in the example of burning the bridge and so do we when we stop in front of red traffic lights. Nonetheless, without giving further details these examples do not live up to the third principle.

According to Blumer "interpretation should not be regarded as a mere automatic application of established meanings but as a formative process in which meanings are used and revised as instruments for the guidance and formation of action" (Blumer 1969: 5). Responses to traffic lights appear to be automatic. While I have shown that game theory can deal with the interpretation in the cases of the traffic light and the burning bridge, these instances might be considered rather simple. Blumer also intends to cover more complex cases by giving more prominence to interpretation processes. More detailed rational choice models also allow the interpretations to play a bigger role.

Consider the case of a driver who sees a red traffic light at a cross-roads ahead at night and it is commonly known that there is very little traffic at that time. The driver can see that no one is driving on the intersecting road. She slows down to make sure she will not overlook anything, but despite the red light she drives through the intersection. The interpretation of the traffic light does not simply lead to an automatic compliance with

the rule. The driver takes the red traffic light as a warning but does not adhere to its injunction because she also considers other information.

We can describe this process of interpretation using rational choice tools. Assume that we have instituted a convention to stop on red and drive on green. But just as Blumer emphasised, the interpretation matters. Let us return to the driver who sees a red traffic light up ahead at night and there is no traffic at all. A rational choice account can model her as facing the following decision tree:[40]



Encounters with a red traffic light split up into two types: encountering the red light during daytime traffic and encountering it alone at night. In each type of situation, the agent faces the decision whether to follow the convention or not. If the agent follows the convention, she stops and receives a pay-off of zero. She does not get to her destination faster, but she also avoids accidents. This holds at night and during the day, but the pay-

[40] It is not quite a game in extensive form since I don't include the other players properly.

off for diverging from the convention differs depending on the time of day and the associated traffic.

If the driver does not follow the convention during the day, she receives a negative pay-off because she is likely to have a crash. During the night, however, ignoring the convention has a small pay-off since it speeds up the journey and there is hardly any traffic.[41] We see that here interpretation is not just an automatic response to the light but takes the environment into account. Rational choice models can capture this perfectly well.

Although this model illustrates that rational choice accounts of signalling leave ample room for flexible principles, we should only consider it a start. For example, I have not introduced probabilities into the decision tree, as a realistic model should. Nothing stops us from adding more complexity to the model and the more we add, the more the agent appears to be a sophisticated interpreter. Standard rational choice theory can account for the complex interpretation processes.

It also allows for such processes to affect conventions and thereby the meaning of objects. Assume that we start out with a convention to stop at red traffic lights. However, it becomes common practice to ignore red traffic lights during the night if no other drivers can be seen. The agents interpret the light more flexibly than originally intended. At some point, this becomes common knowledge and everyone accepts this regularity. Even the police officers won't give you a ticket if they see you crossing a red traffic light at night. Under these circumstances, we have a different convention than before. Accordingly, the meaning of the signal changed because of the way interpretations guide actions. The conveyed information turned from an unqualified "stop" to a conditional one.

[41] I ignore the possibility of law enforcement.

The example illustrates how interpretation can affect meaning. Even if the agent ignores the red traffic light at night, we can still understand the meaning of the traffic light in terms of the information it conveys. Because the incentives in the lonely night-time driving situation differ from those which the convention was supposed to handle, the game changes, which causes the behavioural pattern to change, which in its turn affects the convention. The rational choice models meet Blumer's demand that flexible and intelligent interpretation processes shape meaning.

Preliminary Conclusion

I have considered all three principles Blumer put forward to distinguish symbolic interactionism and found that a rational choice account of signalling meets all of them. Does that mean that standard rational choice has incorporated all theoretical contributions by symbolic interactionism? The assumption of fixed intrinsic preferences limits the role of interpretation in the standard rational choice account. Signals can neither convey the information of intrinsic preference change, nor can interpretations have the force to shift intrinsic preferences.

The next section argues that symbolic interactionism can offer more than standard rational choice accounts of signalling because its pragmatist heritage includes an account of motivational change: the theory of motivationally Problematic Situations. Introducing this type of preference change, information and interpretations can shape agents and their interactions more deeply. Although Blumer himself does not develop this contrast, he follows Mead and emphasises "that human interaction is a positive shaping process in its own right" (Blumer 1969: 66). He continues by describing interaction as follows:

"The participants in it have to build up their respective lines of conduct by constant interpretation of each other's ongoing lines of action. As participants take account of

each other's ongoing acts, they have to arrest, reorganize, or adjust their own intentions, wishes, feelings, and attitudes […]" (ibid.). Standard rational choice models capture more of this process than Blumer knew, but they do not capture everything. Only by giving up on the assumption of fixed preferences can we reconstruct the whole contribution of pragmatist sociologists.

Signalling and Problematic Situations

In his book *Signals*, Brian Skyrms notes that signals do not only convey "*information about the state of nature*, but […] also *information about the act* that will be chosen" (Skyrms 2010: 38, italics in the original). By burning the bridge, the general conveys information about the action that she will choose in the game of chicken: namely, to fight.

Signals can contain information about what might affect the outcomes in a standard rational choice model: states of natures, preferences, chosen acts. The theory of Problematic Situations additionally allows for agents to signal intrinsic preference change and potential new intrinsic preferences. Since intrinsic preference change does not figure in standard rational choice models at all, it cannot figure in standard rational choice models of signalling. Lewis does not discuss preference change in his book and Skyrms concerns himself mostly with evolutionary game theory, which replaces the pay-offs in terms of preferences with pay-offs in terms of evolutionary fitness. Pragmatist sociologists, however, consider the change of goals, including the goals given by intrinsic preferences, to be an important aspect of human agency.

In the previous chapter, I reconstructed the theory of motivationally Problematic Situations as a theory of preference change, including the change of intrinsic preferences. In my reconstruction, a Problematic Situation follows the disruption and confusion characterising an indeterminate situation, and opens the agent up for preference change.

The agent then explores new courses of action in an exploratory phase. This theory allows for the signalling of preference change and for the interpretation of signals to affect preference change, or so I argue in this section.

Other models of preference change, such as the one suggested by Cohen and Axelrod (1984), lack an equivalent to the exploratory phase. But this phase makes all the difference for signalling. The agent can signal within the exploratory phase which preferences she assumed in a qualified tentative mode, that is, without fully committing to them yet. I might signal to my mother that I am reconsidering my distaste for broccoli, but I am not yet fully committed. She might cook broccoli, but I reserve the right to pull back.[42]

For contrast, assume that encountering a Problematic Situation led to an instantaneous change of preferences. In this case the agent might signal that she has undergone preference change and the new preferences she has developed. But since the preference change occurs instantaneously, every response to the change comes after the fact. I can signal to my mother that I now like broccoli, but not that I am reconsidering my distaste. The signals of other agents cannot react to the preference change and affect it.[43]

At the end of the previous chapter, I discussed a case in which the response to preference change comes after the fact: Simon and David want to organise a conference but find themselves in a prisoner's dilemma. To get funding for their conference at least one of them has to read papers while they are on different flights to a meeting with their funding

[42] I actually like broccoli more than my mother does, but I wanted to let the stereotype live on.
[43] Could one not get an equivalent to the exploratory phase if the preference change occurred in small increments? Say one could come to like broccoli a little better, and even more given the appropriate circumstances, and so on. Even though such a process might mimic an exploratory phase, important differences would remain. For example, the exploratory phase allows trying out a different course of action which can only be reached via a large revision of preferences. No small increments would allow us to emulate this.

body. If neither of them reads the papers, they will not receive funding for the conference and if both read them they receive a bit more funding.

They would prefer mutual reading (cooperation) over mutual defection. But both prefer to defect while the other does the reading over both doing the reading. Finally, they prefer mutual defection over doing the reading alone. They face a prisoner's dilemma:

PD		David	
		Read	Defect
Simon	Read	2 (Simon), 2 (David)	0, 3
	Defect	3, 0	1, 1

Since Simon and David are on different flights, they cannot communicate, and so we should expect both to defect. However, as my example goes, Simon encounters a Problematic Situation – he left his USB flash drive back at home and so cannot follow his intention to work on a particular task – and as a result his intrinsic preferences change. Simon becomes more cooperative:

PD		David	
		Read	Defect
Simon	Read	4 (Simon), 2 (David)	2, 3
	Defect	3, 0	1, 1

Simon informs David about his preference change after his flight has landed. But by then he already settled on the new course of action. Simon has read the papers while flying and his signals come after the fact.

Nonetheless, Simon went through an exploratory phase. During the exploration Simon started looking into the papers and found that such a course of action suits him after all. That he settled on these new preferences is a contingent state of affairs stipulated by me. Simon might have found such a course of action unsuitable, even after having had a look at the papers. While encountering a Problematic Situation increases the chances of preference change, according to the pragmatist sociologists, the encounter does not guarantee the change.

If David could interfere while Simon explores, he would try to move Simon towards cooperation. After all, David profits from Simon reading the papers, irrespective of how David himself acts. In Simon's case, the flight rules communication out. All signalling happens after the preference change. Often, however, the exploratory phase occurs in an accessible setting. Then signalling during the exploratory phase plays a significant role.

During the exploratory phase, the agents send and receive signals which convey information about the preference change. The agent in the exploratory phase might send a signal that she is open for new preferences and which preferences are candidates. Simon could signal that he considers ranking reading/cooperation higher. He follows a tentative preference without yet fully committing to the preference.

I introduced the notion of a tentative preference for this purpose in the last chapter. The agent tries out a new choice that is followed from a candidate's tentative preference. Simon looks into the papers and explores how reading them works out for him before he really gets into it. Following the tentative preference, the agent can receive signals that inform her of how her environment perceives the potential preference change. In contrast to standard rational choice theory, symbolic interactionism can account for how signalling affects the development of tentative preferences.

Cooperation can become more fundamental thanks to such signalling.[44] The agents may not only act in mutually beneficent ways, but they might also align their intrinsic preferences as they develop. They exchange information about potential new motivations and thereby make sure the preferences fit together. Consider the following example:

Each day Matilda and Martin go for separate walks after lunch. They can choose between two walking routes: one up the hill and one around the lake. Both Matilda and Martin intrinsically prefer to walk around the lake over walking up the hill and both intrinsically prefer to walk alone rather than together. Martin and Matilda do not particularly like each other. They do not even greet each other. Both prefer to go up the hill if the other takes a walk around the lake. For both, the worst outcome is to meet while walking up the hill. The pay-off matrix looks as follows:

		Martin	
		Lake	Hill
Matilda	Lake	1, 1	3, 2
	Hill	2, 3	0, 0

(The preference for walking alone deducts 2 from every walk the other shares.)

Given this matrix and the repeated nature of the game, Martin and Matilda probably find one or the other routine that allows them to avoid each other. Both lake/hill combinations are Nash equilibria, that is, neither of them can improve the outcome for themselves by diverging from this state while the other sticks to it. [45]

Let us assume that they end up with a convention: one day Martin goes up the hill and Matilda walks around the lake and the next day they do it the other way round. They

[44] I later discuss that such signalling and associated preference change doesn't guarantee a more cooperative outcome. Further empirical assumptions are needed.
[45] Technically we have here an impure coordination problem, also known as Battle of the Sexes, which I discuss in chapter six as a model for climate change negotiations.

maximise their overall pay-offs and cooperate by avoiding each other – until a Problematic Situation occurs.

Martin goes up the hill, but his walk comes to an unexpected halt right around the first corner because a recent thunderstorm has created obstacles. Trees lie across the path so that it has become impassable. Martin enters a Problematic Situation and opens up to preference change. He could go to the lake, but he knows that Matilda is taking a walk there.

Usually the prospect of meeting Matilda would reduce the expected pay-off. It would still be the best remaining option, but with a reduced overall satisfaction. The Problematic Situation, however, opened Martin up to preference change and he starts to think that it might not have to be that way. Maybe he and Matilda could get along after all? He wants to give it a try. While his preferences have not yet changed for good, he enters an exploratory phase where he acts with a tentative neutrality towards sharing the walk with Matilda, rather than his prior aversion.

Martin heads towards the lake and, sure enough, runs into Matilda. At this point signalling comes in. Following their former motivations, Martin and Matilda would ignore each other, but on this occasion Martin greets Matilda and even tries to strike up a conversation while they share the path. His actions convey the information that he might be willing to share the walk with her after all. He signals that his preferences might change. The greeting and attempt to converse increase the probability of a different preference and this is common knowledge between Martin and Matilda.

Matilda can react to these signals in multiple ways. Assume that she keeps to her preference for walking alone. She might respond dismissively to Martin's attempts to talk. If she does this, she signals back that she does not want to share a walk with Martin and that she therefore does not approve if he changes his preferences in such a way.

But Matilda might also enter a Problematic Situation when she is confronted with Martin. Since they usually go separate ways, she did not expect him. Matilda stops walking for a moment and Martin greets her, another unexpected event. He even tries to strike up a conversation! She identifies a problem: the usual division of walking routes has broken down. The Problematic Situation opens Matilda up to preference change.

In her own exploratory phase, she also tries different courses of action, following a motivational profile characterised by tentative preferences. Matilda, too, tries out giving up her preference for walking alone. She signals her exploration to Martin, by greeting him back and joining the conversation tentatively rather than killing it. This reaction contributes to establishing the motivational profile Martin has tried out, and both lose their dislike for sharing a walk.[46]

From now on Martin and Matilda enjoy the walk around the lake together. We have a new game:

		Martin	
		Lake	Hill
Matilda	Lake	3, 3	3, 2
	Hill	2, 3	2, 2

Going to the lake becomes the dominant action for both.

My example sketches a best-case scenario for cooperation. The agents could keep to their old motivations in the situation. Assume that only Martin changed his preference and did not respond to Matilda's dismissive signal. He no longer experiences a decreased pay-off for sharing the walk with Matilda, so that the new game would look as follows:

[46] In the next section I address how this might work. For now, I just assume that signals can have such an impact on the change of intrinsic preferences.

		Martin	
		Lake	Hill
Matilda	Lake	1, 3	3, 2
	Hill	2, 3	0, 2

For Martin, going to the lake dominates. No matter what Matilda does, he prefers going to the lake. This is no surprise, since he only became indifferent towards sharing a walk with Matilda. For Matilda, however, the rational move now becomes to always take the path up the hill, since she can expect Martin to take the dominant action and her pay-off is higher if she goes to the hill while he goes to the lake.

Matilda's pay-off decreases in comparison to the former convention since she never gets to walk around the lake any more. This change on Martin's side is not particularly cooperative. Before at least they cooperated in avoiding each other, now Martin lacks an incentive to do so. The preference change is not coordinated, forcing one agent, Matilda, to live with a smaller pay-off from now on.

But the signalling during the exploratory phase allows them to communicate how they might change their preferences and align in the process. Before the change, Matilda and Martin cooperate in the sense of contributing their share to avoiding each other. During their exploratory phases, they coordinate the change of their intrinsic preferences and both end up with a more satisfying solution. With the aligning preferences, they enter a new game in which both prefer to go to the lake in the presence of each other.

This result is more cooperative than Martin's unilateral preference change, because they now both increase their pay-offs. Martin and Matilda undergo preference change together leading both to choose the lake and experience higher pay-offs than before. Signalling allowed the alignment of the new motivational profiles.

While this example serves as a good illustration for signalling during exploratory phase, it can only be the start of our discussion. At this point it remains unclear how signals can guide the change of intrinsic preferences. Martin might signal his tentative preference to Matilda by greeting her and her response supposedly influences his final motivational profile, but more remains to be said about the underlying mechanism.

Signals, Preference Change, and Cooperation

The example of Matilda and Martin illustrates how agents who only have one shared goal – avoiding each other – give this goal up in an exploratory process coordinated by signalling. But how can the signalling affect the preference change? By greeting her and starting a conversation, Martin sends the signal that he might change his preferences and no longer avoids Matilda. How does Matilda's response, where she greets him back and joins the conversation, influence whether he adopts the different motivational profile he tries out?

In the exploratory phase, the agent tries out tentative preferences and, if the feedback to the trial is appropriate, she adopts new preferences. In the picture offered by pragmatist sociologists, the response of the environment to acting on the tentative preference affects whether and what intrinsic preference change occurs. How other agents react to the new course of action is one response from the environment that affects preference change. The question concerns which mechanism underlies this influence. We can distinguish an inferential and a non-inferential mechanism.

The first option is to postulate second-order preferences guiding the change of preferences. Martin might *prefer* to give up his preference for walking on his own only if Matilda is also willing to give up her preference. This solution suggests a cognitivist picture in which agents consider which preferences they should adopt and then draw the

appropriate inference.[47] On some occasions authors in the tradition of pragmatist sociology have put forward formulations that invite such a reading. Describing Dewey's theory of changing ends, Hans Joas writes:

"Dewey's orientation is not […] towards a blind respect for values and a blinkered pursuit of goals, but rather a pragmatic participation in collective action in which all values and all goals are potential objects of reflection and discussion." (Joas 1996: 155)

The claim that goals change because agents reflect on them suggests that agents infer goals in conscious deliberation. While signalling would allow for such deliberation to become collective, it is not obvious that any deliberation can lead to different intrinsic preferences. As I mentioned in the previous chapter, inference to intrinsic preferences remains controversial.

Furthermore, it is not clear whether pragmatist sociologists endorse such an inference. While I offered a quote from Joas suggesting such an interpretation, Joas himself endorses a picture of agency which reduces the role of deliberation (cf. Joas 1996: 148-167). He emphasises the non-intentional underpinning of action rather than intentional inferences. Dewey himself is not as explicit as Joas' interpretation suggests. We should not be too hasty in attributing to the pragmatist sociologists the position that Martin *infers* a first-order preference based on Matilda's signal.

A reconstruction of the Negotiated Order approach has a second option: a non-inferential psychological mechanism might influence the preference change. Assume that agents are more likely to adopt a new motivational profile after trying it out, if the preferences that remain unchanged are at least as satisfied as before. The agent does not *infer* from being

[47] At least in principle, second-order preferences could also be involved in a non-inferential mechanism. However, if one postulates a non-inferential mechanism, the need to postulate second-order preferences is greatly reduced. I therefore disregard this possibility.

satisfied that she should follow the new motivational profile, rather being satisfied enough *causes* the profile to settle in.[48]

As discussed in the previous chapter, tentative preferences are qualified preferences which the agent holds with less commitment. The continued satisfaction of prior preferences could cause the qualification to fall away and the commitment to increase. By postulating such a non-inferential psychological mechanism, pragmatist sociologists can avoid committing to the claim that agents infer preferences from second-order preferences.

Consider the case of Martin and Matilda again. Martin adopts the new motivational profile because Matilda reacts approvingly when he tries out the new motivational profile. We can interpret the "because" in multiple ways. We can understand Martin to have a second-order preference for preferences that receive approval from other people. This second-order preference leads Martin to adopt the tentative preference after the positive signals from Matilda. In this interpretation, Martin engages in an inference to an intrinsic preference upon receiving Matilda's approving response. The signalling would affect the preference change because it leads to an inference.

However, we can also interpret Martin as having only a first-order preference for approval from other human beings and to postulate a non-inferential mechanism which increases the chance of adopting motivational change if events satisfy this first-order preference. The satisfaction of the first-order preference removes the tentative qualification from the preference Martin tried out. In this interpretation, Martin does not engage in an inference to an intrinsic preference. The signalling would affect the

[48] The non-inferential/inferential should not be confused with sub-agential/agential distinction. At least in principle, agents might be able to engage in sub-agential inferences, that is inferences which are not accessible to agential control. A non-inferential mechanism is one that does not involve inferential justification appropriately.

preference change because it satisfies a first-order preference and this satisfaction affects the change via a non-inferential mechanism.

The pragmatist sociologist can choose from these two options: First, signalling affects preference change via a second-order preference and an inference to a first-order preference. Second, signalling affects preference change via a first-order preference and a non-inferential mechanism.

I am still engaged in an effort to offer a charitable reconstruction of the Negotiated Order approach, and therefore the right option would be the one that fits the literature better. At no point, however, do Dewey, Blumer or Strauss distinguish appropriately between second- and first-order preferences in a way that enable me to just rely on following their word. My reconstruction, therefore, allows for both interpretations. The Negotiated Order approach commits to preference change in response to Problematic Situations and that signalling affects the preference change, but the approach remains open for various mechanisms of how signalling affects the preference change.

My account remains informative, even if I only maintain that signals affect preference change without settling for one mechanism. However, while I can leave the question of the exact psychological mechanism open, both mechanisms I present rely on appropriate preferences supporting cooperation. This reliance raises the question as to which preferences have to be in place for pro-cooperative preference change through signalling.

Consider how I described the case of Martin and Matilda: Martin is more likely to give up the preference for walking alone if Matilda also signals that she approves of the new motivational profile. He tries to strike up a conversation and, because Matilda signals willingness to try walking with him by joining the conversation, the revision of the motivational profile settles in. I assume that Martin's psychological make-up renders him more likely to end up with preferences that align with those of Matilda. On this

assumption, his new profile settles in because Matilda signals approval. But the assumption could be wrong. Martin's new motivational profile could settle in because it displeases Matilda.

Both mechanisms, the one assuming second-order preferences as well as the non-inferential mechanism, could favour anti-cooperative preferences, which lead to rejection by other people. Think of the cliché of a rebellious teenager who develops a preference for a certain kind of music only because this music annoys her parents. After a Problematic Situation, she might give a different band a try during an exploratory phase. Only when her parents signal their disapproval is she hooked. Her preference change is misaligned with the preferences of her parents.

The case of the teenager fits both mechanisms. She could have a second-order preference for preferences her parents disapprove of. Since her parents signal disapproval of the new music, she infers a new first-order preference for it. Or, she could have a first-order preference for disapproval from her parents and the satisfaction of this first-order preference affects via a non-inferential mechanism regarding whether a new motivational profile settles in.

In either case, the result is anything but cooperative. The positive effect of signalling for achieving cooperation depends on the motivations of the agents. The question as to whether signalling during the exploratory phase supports cooperation at the level of preferences becomes empirical. Should we believe that the motivations conducive to the cooperative effects of such signalling are widespread? Within the Negotiated Order approach, authors suggest a motivation that increases the likelihood of cooperation: the agents want to make the group work. The Negotiated Order approach focuses on agents

who want to get something done together (cf. Strauss 1988: 234; 2014: 88). Consider the following example illustrating Strauss' idea.[49]

Henry and Harriet work for a software company and want to get things done in their job. However, the integrated development environment (IDE) they use has become outdated and unable to fulfil its purpose. The inadequacy of the IDE disrupts their activity and they encounter a Problematic Situation. Henry and Harriet have to come to an arrangement about which IDE to use in the future.

Both have intrinsic preferences concerning programming environments, but the Problematic Situation opened these preferences up to change. Henry used to have an intrinsic preference for IDEs that only have basis functionality and that you can extend using plugins, over IDEs that have more functions out of the box but cannot be extended using plugins. Harriet had exactly the reverse preference.

On a standard rational choice approach, we would presume that the preference for getting things done would finally override their intrinsic preferences concerning IDEs. But within the Negotiated Order approach we also have the option that having to settle for a new IDE in response to a Problematic Situation, Harriet and Henry both undergo preference change. They try out various IDEs and after trying one called "NETLentils", Harriet signals that she would like to use it. Henry would usually veto the use of NETLentils because, contrary to his intrinsic preference, it does not support plugins. However, because Henry entered an exploratory phase after the Problematic Situation, he tries it and Harriet's signals contribute to the change of his former preference.

[49] The need to present my own example arises, because Strauss' examples (e.g. Strauss et al. 1963: 164-165, Strauss 2014: 92-93) do not clearly capture the change of *intrinsic* preferences, although he endorses the intrinsic preference change (see previous chapter).

Harriet's signal influences whether Henry settles on the new preference suitable for NETLentils. Again, we can interpret the example as working via a second- or via a first-order preference. Henry could have a second-order preference to develop preferences that help the organisation work over those that do not. Giving up his preference to work with IDEs that support plugins helps the organisation to work, because then he and Harriet can agree on NETLentils as their IDE and get on with the work. He could infer from his second-order preference and the appropriate beliefs a new first-order preference over IDEs.

Or, employing the other mechanism, Henry could have a first-order preference to get work done. The signals from Harriet inform him that they can get work done with NETLentils. The likely satisfaction of the first-order preference then leads via a non-inferential mechanism to the final change of preferences. This description avoids postulating any inference to intrinsic preferences. Both options remain available to Negotiated Order sociologists. In either case, the signalling affects preference change and thereby helps to assure coordination.

In summary, the agents belong to an organisation and have a second- or first-order preference directed at keeping this organisation working. This preference then functions in one way or another as the connecting link allowing signals between the agents to lead to better aligned preferences. This example illustrates the prototypical description of the organisational dynamics by the Negotiated Order approach. The guarantee for the preference to function as a connecting link is not a priori, but rather the practitioners of the Negotiated Order approach have found the assumption of such a motivation helpful for explaining organisational dynamics. I do not independently argue for the assumption, but apply it from a school of sociology.

Using this assumption, the link between signalling and cooperation holds in cases like that of Henry and Harriet. But one might doubt whether signalling during the exploratory phase can help cooperation in cases of prisoner's dilemma.

Signalling and Cooperation: Prisoner's Dilemma

I am especially interested in whether preference change and signalling affecting this change help to overcome a prisoner's dilemma. Many authors think of climate change negotiations as being a case of prisoner's dilemma (e.g. Brennan 2009). While this model proves too simplistic in chapter six, it serves as a good approximation for a worst-case scenario. If climate change negotiations resembled the prisoner's dilemma, we would expect all parties to defect even though they themselves would have profited more from mutual cooperation. Exactly in such a case we would hope that preference change, aligned through signalling, helps us out.

The nasty nature of the prisoner's dilemma, however, also works against signalling. Let us start first with an example assuming fixed preferences. Two agents usually face each other in a coordination problem. They have to meet in one of two places and not miss each other. In response, they have developed a signal for place one and a signal for place two.

Whenever an agent signals place one, they both meet there, and when an agent signals place two they meet there. The agents have common knowledge of the game, that is, they know the incentive structure and that the other signals with the intention of achieving cooperation. We have an example of a coordination problem solved by conventional signalling:

CG	Option 1	Option 2
Option 1	1, 1	0, 0
Option 2	0, 0	1, 1

Assume now that for whatever reason, the game changes and the pay-offs become that of a prisoner's dilemma:

PD	Option 1	Option 2
Option 1	1, 1	-1, 2
Option 2	2, -1	0, 0

What happens to the signalling? Let us assume that the agents play this prisoner's dilemma game only once and they have common knowledge of this fact. Each of them now knows that the other has an incentive to send the signal for one option and then choose the other option. This common knowledge undermines the information formerly conveyed by the signal. The former signal for place one no longer increases the subjective probability that the agent will be at place one.

The meaning of the signalling action changes because of the different structure of the interaction and the common knowledge thereof. The new situation undermines how signalling contributes to achieving cooperation. Even if one prisoner gives the other a thumbs-up from the cell to show that she intends to cooperate, the other prisoner has good reasons to remain wary. The prisoner's dilemma can undermine the support for signalling, especially if it is played for a limited number of rounds.

So far, I have discussed the standard case with fixed preferences, but the troubles carry over to the cases of changing intrinsic preferences. Assume that two agents, A and B, find themselves in a one-off deadlocked negotiation, which we can model as a prisoner's dilemma, but which allows for signalling before the agents make their final choice.

To get a few chapters ahead of ourselves, assume that A and B are group agents. They are states engaged in an arms race with the structure of prisoner's dilemma. Both would prosper more if they mutually disarmed, but each of them has an incentive to defect. We should expect both sides to continue their arms race.

Prisoner's Dilemma		Player B	
		Cooperate	Defect
Player A	Cooperate	2, 2	0, 3
	Defect	3, 0	1, 1

Assume now that both players enter a Problematic Situation and open up to changing their preferences. For example, a natural disaster devastates both states in the arms race. Their arming activities are disrupted and they identify problems. The Problematic Situations open them up for becoming more peaceful, but only if they each believe the other one will become more peaceful too. They would want to signal this to one another so as to coordinate their preference change.

One of the players, A, wants to signal the tentative preference for disarming they consider adopting fully, but which they have not yet adopted. With their new intrinsic preferences, the game would change to the following:

Transformation A		Player B	
		Cooperate	Defect
Player A	Cooperate	4, 2	2, 3
	Defect	3, 0	1, 1

Cooperation becomes dominant for A if they undergo such a preference change. However, they are more likely to undergo such a preference change if agent B convincingly signalled back, that they also might undergo such a change. As long as A has a good reason to believe that B will develop a similar motivational profile, they will themselves settle on

the more cooperative intrinsic preferences. If both underwent the pro-cooperative change it would produce the following matrix:

Transformation (A&B)		Player B	
		Cooperate	Defect
Player A	Cooperate	4, 4	2, 3
	Defect	3, 2	1, 1

They would both prefer cooperation and the game turns into an easily solvable impure coordination problem. The natural way to interpret the sketched scenario is that A shows a tendency to become more cooperative, but only if they believe that B becomes more cooperative as well.

Because the signalling takes place in face of a prisoner's dilemma, however, the signals become untrustworthy. Even if A does not open up to a preference change in favour of more cooperation, they have an incentive to signal such a change if she can hope to lure B into becoming more cooperative. A intends to make B cooperate, while they defect.

On the other side, B also has an incentive to signal back that they might become more cooperative, even if they do not open up to preference change at all. They too can hope to lure the other into becoming more cooperative. The two states in the arms race might hope to fool each other into falling behind. Each of them has an incentive to signal that they consider becoming cooperative, although neither of them really does.

Assuming all this is common knowledge, the situation undermines the informational content otherwise associated with the signals. The states in the arms race cannot trust the other's signals. Just as in the case of fixed preferences, the one-shot prisoner's dilemma undermines how signalling contributes to achieving cooperation. If the climate change negotiations between states took this form, we should expect signalling to fail

during the negotiations. The contribution of the Negotiated Order approach becomes doubtful.

However, not only does the problem carry over from the case of fixed preferences to those described by the Negotiated Order approach, the solution does too. The following circumstances are known to ameliorate the issue:

- The prisoner's dilemma could be recurrent rather than one-shot. In particular, the same agents play repeatedly against each other.
- The signals can be costly.

If the same agents engage with each other repeatedly in a prisoner's dilemma without a definite end, then signals can establish themselves again. If an agent deceives by sending wrong signals and then defecting, others can punish by defecting in the future. Honest signalling receives rewards which allow rational agents to signal (cf. Skyrms 2004: 4-6). The exact outcome depends on the details: How likely is it that agents interact with each other again? Do they know when the repetition ends? Does each game have the same pay-off structure? For a wide range of answers, the agents can overcome the difficulties of the prisoner's dilemma by signalling and then engaging in cooperation.

Costly signals also allow agents to continue signalling in the face of a prisoner's dilemma (cf. Skyrms 2004: 80-81).[50] In the example of the arms race, one state might signal to the other its exploration of becoming more cooperative by shutting down and dismantling an arms factory. If the signal has sufficiently high costs, then faking it does not pay. The whole point of faking it is to get an advantage in the arms race, but dismantling an arms factory undermines the advantage.

[50] The burning bridge case discussed earlier relies on costly signalling.

For a state sincerely tending towards becoming more peaceful, shutting down the arms factory has lower expected costs. After all, achieving cooperation, the state will shut the factory anyway. The incentive for deception is outweighed by the costs of the signal, but the costs are lower for an honest agent.

The problem and the solution pertain to signalling both with fixed and with changing preferences. However, the cooperation achieved with preference change cuts deeper. Even with stable preferences, agents in a prisoner's dilemma can signal cooperation and cooperate if the game is repeated or they develop costly signalling. However, in these fixed-preference cases the cooperation remains fragile.

Assume that the agents managed to cooperate in a repeated prisoner's dilemma because of its repetitive nature. When these agents learn that the next game is the last one, the incentives become once more that of a one-shot prisoner's dilemma. No longer can one punish the defector in the following round. The cooperation breaks down with the new information.

By contrast, assume that the agents in a repeated prisoner's dilemma aligned their preference change with signalling. The signalling was only trustworthy enough to support the mutual change of preferences because the prisoner's dilemma was repeated. But with the preference change the pay-offs in the matrix changed for good:

Prisoner's Dilemma		Player B	
		Cooperate	Defect
Player A	Cooperate	2, 2	0, 3
	Defect	3, 0	1, 1

became

Transformation (A&B)		Player B	
		Cooperate	Defect
Player A	Cooperate	4, 4	2, 3
	Defect	3, 2	1, 1

Even if the agents learned that the next game of the repeated transformed game was the last one, their changed preferences would support cooperation. The cooperation no longer breaks down with the new information. The cooperation is not built upon the threat of punishment for defection, but upon mutual preferences for the cooperative act. This case illustrates that the cooperation cuts deeper with the inclusion of preference change. Accordingly, the theory of Problematic Situations makes a difference to cooperation achieved through signalling.

Chapter Four: Putting the Pieces Together

The previous two chapters introduced the theory of action underlying the Negotiated Order approach. In the second chapter, I discussed the theory of Problematic Situations, as developed by Dewey and taken up by pragmatist sociologists, which offered an account of preference change.

The third chapter turned to symbolic interactionism and its emphasis on meaning and interpretation. I reconstructed a large part of symbolic interactionism using rational choice models of signalling, but also discussed how the pragmatist sociologists make an original contribution. In contrast to standard rational choice models, signalling and interpretation affect intrinsic preference change.

This chapter puts the pieces of my reconstruction together. I present the Negotiated Order approach as an original account of social orders, where preference change in response to Problematic Situations characterises these social orders. Strauss asserted "that social orders are, in some sense, always negotiated orders" (Strauss 1988: 235), but I aim to establish a more modest conclusion. I analyse what it is for a social order to be a Negotiated Order and suggest that at least some social phenomena are helpfully conceived of as exhibiting such an order. The empirical research by Strauss and his team support this suggestion.

First, I introduce the Negotiated Order approach as a theory of organisations. Although the approach does not limit itself to organisations, Anselm Strauss and others first applied it to such cases, for example hospitals. The originality of the Negotiated Order approach becomes apparent by contrasting it with an account of organisations structured around bureaucracy. Instead of bureaucratic rules and hierarchical chains of commands, Strauss emphasised the role of individual agents and the negotiations between them.

In the next step, I provide an analysis of negotiations. Building upon remarks by Strauss, I distinguish negotiations from other forms of interaction such as coercion and manipulation. Negotiations are a form of interaction directed at an agreement, in which the participants seek to address each other's motivations although they face a motivational conflict. Such negotiations can be explicit or implicit. The agents engaging in them might not even conceive of them as negotiations.

Having introduced the notion of negotiation, I provide a full analysis of Negotiated Orders. Negotiated Orders are revealed to be a special type of social orders characterised by Problematic Situations and negotiations. The contributions of the previous two chapters come in at this point. Intrinsic preference change of agents and their signalling belong to the defining features of Negotiated Orders.

The chapter ends by contrasting a standard rational choice model of an organisation facing a problem with the Negotiated Order approach. I draw on Strauss' ideas and illustrate with an example the contribution of the Negotiated Order approach. The consequences are twofold: On the one hand, the Negotiated Order approach allows us to make some predictions about how intrinsic preference change affects social orders. On the other hand, the Negotiated Order approach does not offer specific quantitative models as the standard rational choice theory does.

The Negotiated Order Approach as a Theory of Organisations

Historically, the Negotiated Order approach grew out of applying symbolic interactionism to organisations. Together with his team, Anselm Strauss spearheaded this development with a 1963 study on "The Hospital and Its Negotiated Order". In this text, Strauss et al. report the results of their observation in a psychiatric hospital. They emphasise their

observation that the organisation needs limited and informal agreements to function and that negotiations serve as a salient means for achieving such agreements.

To see the significance of Strauss' contribution, consider an alternative approach: a theory of organisation according to which strict bureaucratic rules, codification, and hierarchical chains of command govern the functioning of organisations. Accounts in the tradition of Max Weber (1990 [1922], for an introduction see Preisendörfer 2011: 97-102) have taken such an approach. The following sketch of such an account serves as a contrast foil for the Negotiated Order approach. I do not claim, however, to offer the best version of a Weberian account, nor that such bureaucracy-centric theories are the only alternative to consider for establishing the Negotiated Order approach as the best theory of organisations. The purpose of this sketch is to present the Negotiated Order approach by giving it an idealised opponent.

That being said, we find an enlightening difference between the Negotiated Order approach and Weberian approaches. For Max Weber, bureaucracy belonged to the iron cage of modernity, which reduced the role of individuality. Official and codified rules guide the interaction of individuals and their own intentions become secondary. A formalised hierarchy provides a chain of command leaving negligible room for negotiations between the levels. Since everything becomes codified, ambiguity falls away, or so a strict bureaucracy in Weber's tradition might suggest. A sophisticated version would grant that some minor role would remain for individual idiosyncrasies and negotiations. However, they would still insist that for explaining how the organisation works, how it settles on its aims and achieves or fails to achieve them, these other factors can typically be ignored.

Given this approach, we would expect that by looking at the official rules of the hospital and the directives issued down a long a chain of command, we would arrive at a

description of how the organisation functions. We might accept some deviations, some minor adjustments resulting from negotiations between individuals, but these exceptions should be mostly random noise.

Strauss and his team questioned such an approach. Based on their observation, they concluded that official rules and commands do not play the leading role suggested by Weberian approaches. They found that “in most sizable establishments, hardly anyone knows all the extant rules, much less exactly what situations they apply to, for whom, and with what sanctions” (Strauss et al. 1963: 151) and “that some rules once promulgated would fall into disuse, or would periodically receive administrative reiteration after the staff had either ignored those rules or forgotten them” (ibid.).

The bureaucratic scheme of the organisations studied only mattered insofar as the agents took account of it in their daily interactions, which happened only occasionally and selectively. The members of the hospital themselves recognised that “too rigid a set of rules would only cause turmoil and affect the hospital’s over-all efficiency” (Strauss et al. 1963: 153). They considered the iron cage of Weberian rationalisation dysfunctional. Strauss and his team agreed with this folk assessment. Rules fail to govern organisational behaviour in a way that satisfies the participants, so the agents only employ the rules when it suits them.

Consequently, Strauss and his team suggested that the interactions of individuals have to receive more attention. A theory of individual agency, rather than of bureaucratic rationalisation from above, became central. Those seeking to explain organisation must consider what the members of the organisation do, rather than to presume it from the official record of rules and command.

Blumer, too, noted that “[i]nstead of accounting for the activity of the organization and its parts in terms of organizational principles or system principles, it [symbolic

interactionism] seeks explanation in the way in which the participants define, interpret, and meet the situations at their respective points" (Blumer 1969: 58). As reconstructed in the previous chapter, agents act based on the meaning objects have for them and engage in complex interpretation processes which reshape the meaning of the objects. Blumer and Strauss hold that these features of interaction are crucial for describing formal organisations.

Strauss and his team found in their research that organisations create Problematic Situations for their members. Decisions and rules have unintended consequences and the members have to find ways of dealing with them. In their foundational hospital study, Strauss et al. describe how an increase of adolescents among the patient population "raised many new problems, and led to feverish negotiative activity" (Strauss et al. 1963: 165). In varying degrees, the change led to unforeseen disruptions in the working routines of the members, who therefore encountered Problematic Situations as described in the second chapter. Not all of these Problematic Situations lead to a change of intrinsic motivations, but they increase the probability of such changes.

These Problematic Situations take place in a context that demands cooperation. The organisation, in this case the hospital, has to continue working, even if the increase of adolescents disrupts the way it functioned before. In the previous chapter I noted Strauss' assumption that the agents in an organisation try to work things out and negotiations are one salient way to do this (cf. Strauss 1988: 234). The bureaucratic rules need time to adapt and the hospital has to keep functioning until they do. The Problematic Situations cause informal negotiations to maintain the daily functioning of the organisation. During such negotiations, signalling can guide the preference change of group members. New preferences are likely to align because of this signalling, as discussed in the previous chapter (remember the example of Henry and Harriet).

In the case of the problems arising from an increased number of adolescents, the negotiations finally ended in a new rule, which provided an upper limit for the number of adolescents to be admitted. But the low-level negotiations on how to deal with the Problematic Situations preceded the resulting rule, which came with the proviso that if the situation required it, the higher levels of administration could look into it again (cf. Strauss et al. 1963: 166). Even the new rule remained up for negotiation, so that if further Problematic Situations were to occur, the rule would have fallen out of favour quickly.

By now we see the rough outlines of the Negotiated Order approach to organisations: Based on his empirical findings, Strauss denies that bureaucratic rules and a chain of command suffice to keep organisations functioning; instead he points to interactions such as negotiations. For describing these interactions Strauss endorses the theory of Problematic Situations, and Blumer's theory of meaning and interpretation. If we combine these elements adequately, we end up with the Negotiated Order approach to organisations.

In the following, I discuss Strauss' notion of negotiations and then present my reconstruction of the Negotiated Order approach along the lines just sketched. This reconstruction of the Negotiated Order approach, however, will not be limited to formal organisations. Although the Negotiated Order approach arose out of research into organisations, it has since been extended. Sociologists have argued that various kinds of groups and social arrangements exhibit a Negotiated Order, including public pools (Scott 2009). Strauss went so far as to assert "that social orders are, in some sense, always negotiated orders" (Strauss 1988: 235).[51] While I commit only to the claim that the Negotiated Order approach can shed light on some important social phenomena, the

[51] As Strauss clarifies in the next sentence, "in some sense" is supposed to guard against the belief that negotiation fully explains all features of the social order.

following reconstruction of negotiations and Negotiated Orders respects the breadth of the approach.

Analysis of Negotiations

Unforeseen difficulties in the hospital, the problems following an increase in the adolescent patient population, lead to negotiations. For Strauss negotiations are a way of "getting things accomplished" (Strauss 1988: 2, 234). Negotiations in the hospital help to keep it functioning in response to such unforeseen complications.

Not all interactions that allow organisations and groups to keep going are negotiations, however. Strauss compares negotiations with manipulation and coercions, which both can serve as means to getting things accomplished (Strauss 1988: ix-x, 221). These forms of interaction serve as a contrast class which highlight the characteristic features of negotiations.

While manipulation, coercion, and negotiations are often mixed in human interaction, they also offer clearly distinct cases. Pure cases of manipulation and coercion serve to meet a goal by bypassing the motivations of the other agent. If I tell you the wrong date of the departmental meeting, I manipulated you into missing it by supplying false information. If I tie you down with chains and shackles, I coerce you into missing the departmental meeting. If I offer you an incentive not to attend the departmental meeting, I negotiate with you.

To negotiate with another agent implies an attempt to reach agreement based on their own motivations. Nonetheless, negotiations do not imply altruism on any side of the interaction. Two agents haggling on the market negotiate without being altruistic at all. Neither do negotiations imply that one tries to meet all motivations of the other agent,

one only has to try to meet enough of them to reach one's goal. The hagglers in the market try to meet each other's demands so that the exchange occurs, but no more than that.

Of course, a negotiation might also fail because the agents are, in the end, unwilling to meet each other's demands. The two hagglers might be unable to agree on a price, but as long as they try to find a price that suits the motivations of both, they negotiate. If one of them robs the other, the interaction turns predominantly into coercion. If the haggler knowingly pays with counterfeit money, she engages in manipulation.

For many interactions the negotiative, coercive, and manipulative elements are difficult to untangle. When the boss in the company negotiates with the personnel about how they will divide up work this week, the boss can use coercion.[52] She can simply assign one employee to one slot regardless of his own preferences. When the boss splits up the work for this week and informs everyone that they have to do this or get fired, she tries to circumvent some motivations of the employees. She bypasses their desires for a certain division of work and rather coerces them. But, of course, negotiative elements remain as well, since the boss is still taking into account the employees' motivation to keep their job.

By contrast, if she considers their motivations and offers incentives for accepting the division of working time, she engages more directly in negotiations. How much the agents, in this case the boss, use their power to limit the amount of motivations that can even be met in the interaction determines the degree to which the interaction is coercive rather than negotiative.

[52] "Coercion" as used by myself and Strauss does not imply illegitimate force. A police officer correctly arresting a criminal would count as coercion based on my and Strauss' use.

Given these boundaries suggested by Strauss' discussion of coercion, manipulation and negotiation, I propose the following analysis:

An interaction between two or more agents is a negotiation if and only if

i. the interaction is directed at establishing and/or maintaining an agreement,
ii. the agents have to overcome a motivational difference for the agreement,
iii. the agents seek to address motivations of one another in this interaction in overcoming this difference, and
iv. the agents suppose that the matter to be agreed upon depends on their agreement.

For an interaction to be directed at an agreement, the direct participants in the interaction need to have a motivation to agree and this motivation must guide the interaction at least partially (condition 1). If a boss and her employees have no motivation to reach an agreement on how to split up the work, then they do not negotiate. The boss might still try to coerce the employees into working, but she is not motivated to reach an agreement.

Strauss and his team found that in actual organisations the wish to keep the group working usually motivates group members to look for an agreement. They need to agree on one course of action or otherwise the hospital, the corporation, or the family cannot function. In standard cases, the pure imposition of an order from above does not allow a group to achieve its goals. Groups rely on their members to try to achieve agreements.

Agreements can concern a variety of things, such as how to divide working hours or the basic goals shared by a group.[53] When the personnel at the hospital starts to negotiate

[53] I rely on an everyday understanding of agreement. Gilbert 1993 and Mintoff 2004 have proposed two accounts. As far as I can see both should suit my purposes, although Mintoff's account restricts agreements more by analysing them as exchanges of intentions.

about the problems caused by the increased adolescent population, they might find middle ground by establishing a rule limiting the admission of adolescents. They have come to an agreement about how to handle the situation.

The first condition distinguishes real negotiations from pretence negotiations. In pretence negotiations, one of the parties only acts as if they are trying to come to an agreement with the other, while secretly aiming for something else. During a war, a party might pretend to participate in peace negotiations, while in fact not trying to come to an agreement with the other parties, but rather using the time to regroup. The interaction might be considered a case of manipulation rather than negotiation. For a negotiation to take place, all participating agents must aim for an agreement. Such pretended negotiations illustrate that an agent can be mistaken about whether they are engaging in negotiation with another agent. The other party might believe that they are engaging in peace negotiations, while the interactions are only pretence negotiations.

For the interaction to be a negotiation, the agents need to face motivational differences that need to be overcome to arrive at an agreement (condition ii). This motivational difference exists if and only if the motivations of the agents do not allow an agreement without trade-offs. There has to be some conflict between the desires of the agents that is required to overcome in order to make agreement possible.

For example, two employees want to take the Thursday morning shift, but only one slot is available. These two employees have a motivational difference insofar as their immediate desires are not co-realisable. Both desire a scarce resource, which they cannot share. But such a motivational difference does not rule out an agreement, since the parties can engage in side transfers. They might negotiate an exchange of other slots until they find a solution to which both can agree.

The demand for a motivational difference rules out cases in which one agent makes a proposal and the other happily agrees right away. One employee proposes to another that she will do the Monday shift and the other one will do the Tuesday shift. The other accepts, because this suits her motivational profile as well. In this case the agents are engaging in an interaction directed at an agreement, but they do not have to overcome a motivational difference. The case fulfils all conditions for a negotiation except for the second. The two employees have reached an agreement without having had to engage in any negotiations. For one agent to make an offer and for the other to accept it is an interaction, but it is not enough for a negotiation. We don't negotiate every time I offer you a biscuit and you accept it.

Including such a motivational difference in the conditions not only suits our everyday use of the term, but also Strauss' sociological use. According to Strauss, negotiations serve to work things out. If everyone can agree to everything immediately, they have nothing to work out. We do not work things out when I offer you a biscuit and you accept it. However, if our desires conflict because both of us want the last biscuit, we start to discuss the situation, and you promise to do the dishes in exchange for the biscuit, then we have a negotiation.

The third condition, that agents seek to address the motivations of one another, distinguishes negotiations from pure coercion and manipulation. These other forms of interaction try to bypass the motivation of the other agent. Mixed cases occur if agents try to bypass some motivations and meet others. If the boss uses her power to bypass the employee's preference not to work on Friday, then she coerces her employee. But the boss might still negotiate as to when the employee works during the rest of the week. As a result, the entire interaction is a mix of negotiation and coercion.

This condition demands that *all* participating agents seek to meet the motivations of one another. If in an interaction between two agents only one tries to meet the motivation of the other, this is not a negotiation. Take the example of a kidnapping where the hostage tries to convince the kidnapper to let him go. He makes promises to her on the condition that she lets him go, but the kidnapper just silences the hostage. In this scenario, the hostage tried to engage in a negotiation, but the attempt failed. No negotiations took place.

Accordingly, an interaction among three agents is only a negotiation among all three of them if each of them tries to meet the motivations of the others. If only two out of three do so, then only the interaction between the two counts as a negotiation.

The fourth and last condition demands that the agents suppose that the negotiated matter depends upon their agreement. An appropriate account of agreement might already provide such a limitation, but since I provide no account, but rather depend on our everyday understanding, I included the clarification in my analysis of negotiation.

The condition rules out that physicists engaging in a discussion about whether a law of physics holds are thereby negotiating the laws of physics. The physicists engage in an interaction directed at a theoretical agreement and they try to meet each other's motivations. Nonetheless, they do not negotiate the laws of physics. Negotiators can only negotiate what they deem to be within their power. The physicists can negotiate which model to use in a simulation, but they cannot negotiate the laws of physics.[54] My definition allows for mistaken negotiations, that is, cases in which the agents mistakenly believe that the matter is up to them.

[54] I assume here that they endorse a sort of realism about the laws of physics. According to such realism one might negotiate how to formulate a physical law, but the law remains independent of the formulation. An analogous case can be constructed for moral realists, who discuss what they morally ought to do.

Those are the four conditions for negotiations. One might be tempted to add a fifth condition demanding common knowledge of the previous four, i.e. that each agent believes that the conditions are fulfilled, and that the other believes that they are fulfilled, and so on. I do not see the need for this further limitation. Consider the case of the two warring parties engaged in peace negotiations, in which each side believes that it is more likely than not that the other is just using the interaction to gain time to regroup. But while both have this fear and therefore lack the common knowledge that the first condition is fulfilled, both in fact meet all the conditions and try to reach an agreement in their interaction. I would call this a case of a negotiation. Therefore, we should not include common knowledge among the requirements for negotiations.

My proposed analysis of "negotiation" remains very broad, but this fits Strauss' use. In particular, Strauss asserts that "negotiations can also be implicit, their products being tacit agreements or understandings" (Strauss 1988: 224). Consider the following example.

A boss and her employees want an unspoken agreement about how many minutes late one can arrive in the morning. Given their different roles, they face a motivational difference that needs to be overcome: the boss wants the employee to be on time as much as possible and the employee wants to have some leeway. Even without words spoken, we can find side transactions: one employee might stay longer, hoping that the boss in return will become more lenient about arriving late. They try to meet each other's motivations and they suppose that the matter to be agreed upon is up to them. They negotiate implicitly about how late employees can come.

Negotiators do not have to explicitly state that they are engaging in negotiations. They might even be surprised to learn that they engaged in negotiations. Nonetheless they can meet the necessary and jointly sufficient condition for a negotiation.

While not an essential part of the analysis of negotiations, they notably serve as means to address Problematic Situations within organisations. During such negotiations, participating agents have a propensity to change their intrinsic preferences and signal so to each other. The previous chapter's example of Harriet and Henry's negotiation about which integrated development environment (IDE) to use illustrates this connection.

Harriet and Henry try to settle on one program for the software company to work. Their interaction is directed at agreement and they must overcome a motivational difference. They intrinsically prefer different types of IDEs. Neither of them tries to bypass the motivations of the other, but rather they seek an agreement suiting everyone. In other words, they negotiated with one another. However, they overcome their motivational difference not just by engaging in an exchange. Henry undergoes a change of intrinsic preferences that allows them to leave the motivational difference behind.

In the Negotiated Order approach, agents can overcome the motivational difference necessary for negotiations by a change of intrinsic preferences. A standard rational choice perspective on negotiations neglects this option.

Analysis of Negotiated Order

With my analysis of negotiations in place, I can reconstruct the core of the Negotiated Order approach: the notion of a Negotiated Order. While Strauss introduced this technical term, which I capitalise to distinguish it from the colloquial use, he never offered an explicit analysis in terms of necessary and jointly sufficient conditions.

It is a conceptually necessary feature that Negotiated Orders are social orders which cannot be explained by a model of strict bureaucracy, but are shaped by Problematic Situations and reactions to such situations, including negotiations and intrinsic

preference change. All elements from the previous chapters come together in this notion of the Negotiated Order:

A social order is a Negotiated Order if and only if

(i) its features cannot be satisfactorily explained by reference to bureaucratic rules,
(ii) agents within the social order encounter Problematic Situations threatening the persistence of the social order,
(iii) these situations cause these agents to negotiate with one another, and
(iv) the resulting negotiations shape the social order,
(v) the Problematic Situation opens agents up for preference change, and
(vi) these agents signal with each other during the exploratory phase.

Strauss does not analyse the term "social order" further. As an approximation, a social order is a sustained pattern of a social phenomenon. In this sense, an organisation exhibits a social order. If problems arise in the organisation, the members follow certain procedures and look to certain other members for help. Paradigmatically, families, corporations, and states exhibit such social orders. For example, in a family the father might always do the dishes, in the corporation the boss might fire employees, and in states the head of government might be elected. All these examples illustrate parts of the social order.

Negotiated Orders are social orders that fulfil the six conditions given above. For example, the pattern according to which the hospital investigated by Strauss and his team functions, is a Negotiated Order because it fulfils all six conditions.

The first condition denies that the features of the social order can be satisfactorily accounted for in terms of bureaucratic rules. This is a conceptual feature of Negotiated Orders rather than just an empirical finding. As we have seen at the beginning of this chapter, Strauss and others developed the Negotiated Order approach in contrast to

theories that explain the features of organisations by pointing to bureaucratic rules and the associated hierarchic chains of commands. Given the importance of this contrast for the Negotiated Order approach, I included the denial as condition (i).

Instead of bureaucratic rules and commands, the Negotiated Order approach emphasises informal agreements. In the psychiatric hospital they investigated, Strauss and his team found the "somatically oriented physicians have long-standing agreements with a secretary who is attached to the two wards upon which their patients tend to be housed" (Strauss et al. 1963: 162). We find a pattern of interaction between certain physicians and a certain secretary associated with a pattern of housing patients. The pattern results from an informal agreement rather than a codified bureaucratic rule.

According to Strauss et al. no one knows the hospital "on any given day unless he has a comprehensive grasp of what combination of rules and policies, along with agreements, understandings, pacts, contracts, and other working arrangements, currently obtains. In any pragmatic sense, this is the hospital at the moment: this is its social order" (Strauss et al. 1963: 165).[55] Only knowing the bureaucratic rules and the commands along a hierarchical chain does not suffice for explaining the social order. Therefore, the social order of the hospital fulfils the first condition.

While this first condition is directed against Weberian theories of bureaucratic organisations, it allows that bureaucratic rules to play a role in explaining the features or the social order. Codified safety rules ordered from above make a difference in the psychiatric hospital. However, if we tried to account for the order only in terms of bureaucratic rules, the explanation would remain partial and miss important features of the social order. The other conditions for a Negotiated Order specify further factors that

[55] Strauss and his team leave aside the coercive and manipulative elements of the social order in this quote, which were probably substantial in a psychiatric hospital during the early 60s.

must be included in a satisfactory explanation and should be familiar from the previous discussions.

Condition (ii) introduces Problematic Situations. The agents within the social order encounter a Problematic Situation that threatens the continuance of the social order. In the case of organisations and other groups, the agents are the members. To refer to the example in the psychiatric hospital given above, an increase in the number of adolescents among the patient population disrupted the action routines of the staff. The hospital's patterns of functioning were under stress and the members could no longer follow their usual routines. A Problematic Situation occurred.

As indicated by the capitalisation, I use "Problematic Situation" in the technical sense introduced in chapter two. According to this sense, a Problematic Situation arises out of the disruption and confusion characterising an indeterminate situation, starts with the identification of a problem, and causes the agent to open up to preference change. The change of intrinsic preferences is not ensured, but more likely. All of this applies in the case of the hospital. The higher number of adolescents leads to disrupted activity and confusion, the members identify the problem, and they open up to preference change. Admittedly, Strauss and his team do not explicitly distinguish between intrinsic and extrinsic preferences, but given the background of their pragmatist action theory, I suggest an interpretation that includes intrinsic preference change as the most charitable one.

Condition (iii) specifies that agents used negotiations to cope with these Problematic Situations. As Strauss and his team write, the problems resulting from the higher number of adolescents "led to feverish negotiative activity" (Strauss et al. 1963: 165). These negotiations give the Negotiated Order its name and thus their inclusion in the necessary conditions shouldn't come as a surprise.

Condition (iv) adds that the negotiations reshape the social order. If the negotiations always failed, so that no agreement was reached, they might leave the patterns of the group's functioning untouched. But in this case the negotiations would remain too peripheral to warrant calling the social order a Negotiated Order. The negotiations must reshape the social order. Usually they will do so by establishing new agreements, which include agreements to rescind former agreements.

In principle, failing negotiations, those which do not end in an agreement, can also change the patterns of interaction in a group and therefore the social order would meet the conditions for a Negotiated Order. For example, the group members might start to dislike each other because of the failed negotiations, and therefore stop cooperating. My conditions for a Negotiated Order would be met by such a case. However, the core cases discussed in the literature concern successful negotiations.

This fourth condition has a basis in Strauss, since he and his team wrote that renegotiations in response to Problematic Situations lead to "consequent changes in the organizational order" (Strauss et al. 1963: 165). In the case of the hospital, the negotiations result in a new rule that influences further admissions of adolescent patients.

Condition (v) specifies that the agents have to function according the theory of Problematic Situations and open up to preference change.[56] As discussed in the second chapter, agents enter an exploratory phase after opening up to preference change during which they explore different courses of action. This exploratory phase will typically occur together with the negotiations prompted by the Problematic Situation.

[56] Strictly speaking, condition (v) is redundant since condition (ii) already employs the technical notion of a Problematic Situation and this notion implies opening up to preference change. However, to clarify this issue and underline the importance of intrinsic preference change I added this condition.

Consider the fictional case in which the increasing number of adolescents leads to a Problematic Situation and the personnel opens up to preference change. In response, one psychiatrist might become more likely to give up her intrinsic preference which underlies her preference to work on Tuesdays rather than Mondays. She might explore these new intrinsic preferences in the exploratory phase coinciding with the negotiations.

The last condition, (vi), refers to this exploratory phase and demands the agents signal in this phase. In chapter three I offered an account of signalling affecting preference change. The psychiatrist trying out new intrinsic preferences leading to her indifference regarding working on Monday or Tuesday signals so to other participants of the negotiation. As I argued in the last chapter, the response from other participants can then increase the chance that this preference change occurs. That way, a deep form of cooperation as described in the previous chapter becomes possible.

All social orders that fulfil these six conditions are Negotiated Orders. However, one might object that not all social orders, which agents negotiate, fulfil these conditions. At least conceptually, a population of agents with fixed intrinsic preferences can negotiate with one another. These negotiations can shape the social order of the agents with fixed preferences. We find an order which is negotiated and does not fit the analysis, because conditions two and four are not met.

At the root of this objection lies a confusion invited by Strauss' terminology. "Negotiated Order" is a technical term, as I indicate by capitalisation, not the result of predicating "being negotiated" to "order". Not all social orders, which agents negotiate, are Negotiated Orders. Just as in the case of the term "Problematic Situation", we have to distinguish a colloquial and a technical use. My analysis provides the necessary and jointly sufficient conditions for the technical term "Negotiated Order" as employed by Strauss and others who follow his approach.

I provide necessary and jointly sufficient conditions for a social order to be a Negotiated Order, but one can also say of a social order that it is a Negotiated Order to a greater or lesser degree. One hospital might exhibit more of a Negotiated Order than another. I propose that two variables determine the degree:

1. the extent to which negotiations in response to Problematic Situations shape the social order,
2. the extent to which agents function according to the theory of Problematic Situations and signalling.

If a social order creates few Problematic Situations, or the agents rarely deal with these Problematic Situations by negotiating, then the social order is a Negotiated Order to a lesser degree. For example, if the agents deal with most problems by coercing one another, then the order is hardly a Negotiated Order.

In the Negotiated Order approach, sociologists usually concern themselves with social orders that are Negotiated Orders to a high degree. If coercion dominates a social order, the phenomenon is of little interest to Negotiated Order sociologists. However, the assumption of the approach is that very few social orders remain Negotiated Orders to a small degree. Even all-out war exhibits a Negotiated Order to a considerable degree. Although the parties engage in coercion with one another, internally they have to negotiate. Strauss clearly thought that for most if not all social orders, the Negotiated Order perspective proves illuminating.

In the next section, I discuss what difference it makes for understanding an organisation if it has a Negotiated Order rather than a social order fitting the standard rational choice approach.

Comparing the Standard Rational Choice Theory with the Negotiated Order Approach.

I introduced Negotiated Orders by contrasting them with a quasi-Weberian approach to organisations emphasising bureaucracy. Strauss and his followers stressed the role of daily interactions and, in particular, negotiations rather than those of codified rules and chains of command. But while this contrast allowed me to emphasise the characteristic features of the Negotiated Order approach, the rational choice approach is not committed to a quasi-Weberian take on social orders. Common game theory models also apply to informal negotiations just as much as to bureaucratic structures. As discussed in previous chapters, however, the Negotiated Order approach introduces intrinsic preference change and therefore differs from standard rational choice theory. In this section, I develop an example to illustrate the difference.

Since in the end standard rational choice theory is the rival to my proposal, I have to show the interesting differences that result from introducing intrinsic preference change as suggested by the Negotiated Order approach. In this section, I develop an example for this purpose that takes inspiration from Strauss et al.'s discussion of how an increase in the adolescent population disrupted the hospital. I first describe the case assuming stable preference before I introduce preference change in response to Problematic Situations.

Assume that two nurses, let's call them Bert and Andrea, engage in a daily programme with their patients on the ward. For the programme, Andrea and Bert have to split the patients into a group of all adolescents and a group of other patients. Each of the two nurses then carries out the programme with one of the two groups.

I stipulate, furthermore, that Bert and Andrea differ in how much effort it costs them to carry out the programme. Bert can deal twice as well with adolescents than with non-adolescents. He is indifferent as to whether two adolescents or one non-adolescent are

added to his group. Assuming that we can provide a utility function for Bert, we can specify the indifference relation as follows: 2*u(added adolescent) = u(added non-adolescent). The utility is negative since the larger the group the more effort is required. For convenience, let us assign the following utility values: u(added adolescent) = -1; u(added non-adolescent) = -2.[57]

Andrea has exactly the reverse preference. Adolescents cost her more effort than non-adolescents. She is indifferent regarding the options of adding two non-adolescents or one adolescent to her group. Thus, for her we can write the following equation: 2*u(added non-adolescent) = u(added adolescent). Let us assume that the values are exactly the reverse as for Bert: u(added adolescent) = -2; u(added non-adolescent) = -1.

Before the relative rise in the adolescent population, the ward has 10 adolescents and 10 non-adolescents. Andrea and Bert can run the programme with each of the two groups. If both groups undergo the programme, then they both receive the utility of having their work accomplished. I stipulate the following utility value: u(both groups undergo program) = 15.

If Bert works with the adolescents and Andrea works with the non-adolescents, Bert experiences a negative utility of -10 from the work, but a positive utility of 15 from having it accomplished. His overall payoff would be 5. Likewise, Andrea would receive -10 utility points and 15, so that she would experience the same overall payoff as Bert. If both carry out the programme with the same group, then both experience a high negative utility since they do not receive the positive utility of having their work accomplished, but have to spend the effort on the one group.[58]

[57] For simplicity, I assume there that the marginal (negative) utility remains the same.
[58] I assume that the costs of their work are not reduced if they both undertake the programme with the same group. This might be the case, because they only experience costs by virtue of having to spend time with the members of the group.

Going through the calculations for all options, we can specify the game Andrea and Bert face:

		Andrea	
		Adolescents	Non-Adolescents
Bert	A	-10 (Bert), -20 (Andrea)	5, 5
	NA	-5, -5	-20, -10

(I write “Adolescents” or “A” for the choice of undertaking the program with the group of adolescents and “Non-Adolescents” or “NA” for undertaking it with the group of non-adolescents.)

Both ways of dividing the groups between Bert and Andrea, so that they both work with one of them, are Nash equilibria. In either case, the agent has no incentive to change the action if the other agent sticks to her action. However, one of those equilibria has a better payoff for both nurses. Since Andrea and Bert can communicate, they find it easy to settle on this combination of choices. Accordingly, they follow a daily pattern of action: Andrea implements the programme for the non-adolescents and Bert for the adolescents. Both receive a positive utility of 5 during each round.

This division of labour is part of the social order of the hospital. It belongs to the pattern of interaction that keeps the hospital functioning. So far, negotiations play no role for shaping the social order.[59] Bert and Andrea do not have any motivational differences to overcome. They simply agree to divide the two groups to maximise their utility.

[59] The interaction to agree on one Nash equilibrium does not constitute a negotiation because the motivational difference is lacking. They both want to settle on the one equilibrium solution.

At some point, however, the hospital admits more adolescent patients so that the overall ward size increases and the proportion of adolescents and non-adolescents changes. Before, the population on Bert and Andrea's ward was split 10:10. Now the population is split 25:3. They have a group of 25 adolescents and a group of 3 non-adolescents. Assuming the same utilities given above, the game changes to the following one:

		Andrea	
		Adolescents	Non-Adolescents
Bert	A	-25 (Bert), -50 (Andrea)	-10, 12
	NA	9, -35	-6, -3

No intrinsic preferences have changed, only the extrinsic preferences following from intrinsic preferences have changed because of the different situation the two nurses face. Nonetheless, we find an importantly different game. The common-sense solution of Bert carrying out the programme with the adolescents and Andrea with the other group is no longer an option. Since the group of adolescents is now much larger, Bert would also prefer to work with the non-adolescents while Andrea works with the adolescents. Andrea, of course, prefers to stick with the old pattern of splitting up the groups.

Andrea and Bert are headed for a conflict. They face a motivational difference and so the two nurses start negotiating. Mutual defection, in this case both Bert and Andrea treating the non-adolescents, satisfies neither of the two. They look for an agreement which satisfies both sides.

As a one-shot game the above matrix provides little hope for reconciliation. However, ways out of the conflict remain on a standard rational choice account, because the situation consists of more than the one matrix. First, the game is repeated which allows

for compromises over multiple rounds. Bert and Andrea could endorse a mixed strategy. For two rounds Bert takes the adolescents and Andrea the non-adolescents, then Andrea takes the adolescents for one round and Bert the non-adolescents. The average payoff of this mixed strategy is 5/3 for both.

Second, they could engage in side-transfers. Andrea might propose to take over an unpleasant shift from Bert in exchange for sticking to the old pattern. If Andrea can compensate for the difference of utility, Bert will accept the exchange.

Or Andrea and Bert could lobby for a limitation on the overall number of adolescents, so that they can return to their former more cooperative game. This fits with Strauss' observation that the negotiations ultimately led to a cap on the adolescent population. How the cap affects the game depends on how strict it is.

Bert and Andrea can combine different solutions, for example a mild cap on adolescents with a mixed strategy. In all these scenarios, Bert and Andrea negotiate and this interaction shapes the social order of the hospital. Whether the nurses agree on a mixed strategy or agree to lobby for a cap on adolescents, or both, their negotiations shape the pattern according to which the hospital functions. However, while the social order of the hospital is negotiated in these cases, it does not become a Negotiated Order in the technical sense.

I assumed fixed intrinsic preferences underlying a changing game. But the Negotiated Order approach allows intrinsic preference change in response to Problematic Situations. It demands that agents open up to preference change. To contrast the Negotiated Order approach with the standard rational choice models, I provide an example of how the inclusion of intrinsic preference change can affect the change of the social order.

We start with the same situation, the same agents, the same preferences, the same game. In the beginning, Andrea and Bert have no problem settling on one solution: Andrea works with the non-adolescents and Bert with the adolescents. However, again more adolescents join the ward. We go from a ratio of 10:10 to the 25:3 ratio.

The changed number of adolescents leads to a disruption of prior activity, of the routines the nurses followed before, and confusion ensues. Bert and Andrea identify a problem. Upon encountering the Problematic Situation resulting from an increased adolescent population, Bert opens up to preference change. I gave the following utilities for adding another patient: u(added adolescent) = -1; u(added non-adolescent) = -2.[60] As we have seen, this leads to a game with conflicting motivations. But that was before we introduced preference change in response to a Problematic Situation.

Assume now that in response to the Problematic Situation, Bert and Andrea start to negotiate. During these negotiations Bert tries working with more adolescents. He has a new set of preferences favouring adolescents even more. He signals the new course of action to Andrea who signals back approvingly. (She has a reason to do so as we will see.) Bert finds the tentative motivational profile he explores congenial and his preferences change accordingly. Adolescents are even less costly for him. Let us assume that u(added adolescent) becomes -0.5 for Bert. On this assumption, we also have a new game, but a different one than before:

		Andrea	
		Adolescents	Non-Adolescents
Bert	A	-12.5(Bert), -50 (Andrea)	2.5, 12
	NA	5, -35	-6, -3

[60] For simplicity, I assume again that the utility does not marginally decrease.

Bert and Andrea are no longer heading for a conflict. The same solution as before, Bert taking care of the adolescents and Andrea the non-adolescents, stands out as the only Nash equilibrium.

Andrea and Bert can overcome the difficulties resulting from the increase in adolescents not only by finding a compromise mixed strategy, side-transfers, or imposing a cap on adolescents. A change of preferences can also help to maintain an old routine. We see that including preference change opens the space of potential developments the social order can undergo in response to unforeseen events.

These results should not come as a big surprise. No one questioned that with different preferences, agents react to situations differently. The finding that preference change influences the social order of groups is important but expected.

The significant contribution of the Negotiated Order approach is not so much to introduce change of intrinsic preferences in general, but to give a specific framework for how these changes occur. They occur in response to Problematic Situations and are part of an exploratory phase during which the agents can signal.

Bert and Andrea do not just randomly change their intrinsic preferences, but they have an increased chance of doing so in response to Problematic Situations. Furthermore, they can signal and therefore align in their preference change. Accordingly, we have a higher expectation that they change preferences in a way that maintains the basic functioning of the organisation. The framework of the Negotiated Order approach gives us clues about the consequences of the intrinsic preference change.

Admittedly, endorsing the Negotiated Order approach renders predictions more difficult. Assuming fixed preferences, the information about the changed number of adolescents

allows us to predict in which game Andrea and Bert will find themselves. But if Bert and Andrea might undergo some preference change with some thus far unspecified probability, we have hardly a clue what game results from the Problematic Situation.

Before the adolescents join the ward, we can only lay out a variety of games that might result from the increase. The current Negotiated Order approach lacks the means for a detailed prediction. A group with a Negotiated Order becomes harder for us and for its members to predict, but that does not prove that the account is wrong or uninformative; it provides guidance for limited predictions. Furthermore, by including preference change it also opens possibilities. What seems doomed from the perspective of standard rational choice theory, can appear more hopeful in light of potential preference change. This difference matters in my later discussion of climate negotiations. First, however, I turn to the question of group agency.

Chapter Five: Group Agency for the Negotiated Order Approach

After having reconstructed the Negotiated Order approach, I now turn to the question of group agency. After all, I intend to apply this theory to the interaction of group agents, paradigmatically to climate change negotiations. In this picture, states serve as negotiators and at least some of them follow the theory of motivational change described by the Negotiated Order approach and reconstructed by me.

As is common in the debate on group agency (e.g. List & Pettit 2011, Huebner 2014, Tollefsen 2015), I will use the term "group" broadly so as to include organisations and states. Group agents are groups which have and act on propositional attitudes such as beliefs, desires, and intentions. For example, a philosophy department that intends to hire a philosopher in a popular field and believes that metaphysics is a popular field, and that accordingly hires a metaphysician, is a group agent.

Robert Wilson (2001) has distinguished the "group mind" thesis, which attributes mental capacities to groups, from the "social manifestation" thesis, which attributes special mental capacities to individuals in social contexts. For example, Searle's (1990) account of joint action attributes special mental capacities to individuals, namely the capacity to have intentions in a primitive we-mode, but denies the group mind thesis. Postulating group agents, such as a philosophy department intending to hire a professor, clearly implies the group mind thesis.[61] In our case, the idea is not that the individual negotiators in climate change negotiations exhibit special mental capacities in this particular social context, but rather that the states are the negotiators.

[61] For a discussion of the difference between joint action and group agency see also Pettit & Schweikard 2006.

As Wilson noted, the group mind thesis is the more controversial of the two. It creates a number of difficult problems in social ontology and philosophy of mind. However, within the constraints of my thesis I cannot address all issues associated with group agency. Instead I provide a map of these issues and pick out those that can be answered within my project.

Overall, we can distinguish five salient questions that arise from taking climate negotiations between states to be an interaction of group agents. First comes the question of which groups, if any, are agents. This formulation encompasses the controversial issue regarding whether group agents exist at all (for discussion see French 1979, Rupert 2005, 2011, List & Pettit 2011, Pettit 2014, Huebner 2014, Tollefsen 2015, Epstein 2015, 2017, Ludwig 2017). While one of the most prominent debates within philosophy of the social sciences and social ontology, I will avoid this issue because it is settled by the school of sociology I follow within this thesis.

Strauss asserts that group agents exist.[62] Already when he introduces the term "actor" he includes groups, writing that an actor "will be the agent of an action–a person, a group, an organization, or other social unit" (Strauss 2014: 23).[63] Strauss' (1988: 210-218) example of the Balkan negotiations between the USA and the USSR can be interpreted as an illustration of negotiations between group agents. He explicitly intended it to undermine the view that the Negotiated Order approach is limited to the microsphere of individuals (see Strauss 1988: 249). In sum, there remains little doubt that Negotiated Order sociologists attribute agency to groups,[64] despite the fact that they tend to focus

[62] Herbert Blumer also seems to allow for group agents. However, he endorses a picture of group agents where a dictator or an oligarchy directs the action of the group, making it little more than the extension of individual agency (see Herbert Blumer 1969: 55-56).
[63] I assume here and in the following that organisations are a special type of group. This fits with debate on group agency (see the examples of List and Pettit 2011 and Tollefsen 2015).
[64] For one example of a Negotiated Order sociology paper that attributes agency to organisations, see Nathan & Mitroff 1991.

on individuals. I rely on this sociological endorsement of group agency rather than arguing for it independently.

A second and closely related question, however, is not addressed by Strauss: Which theory of mentality, that is of mental states, should we employ for groups? While I can rely on pragmatist sociologists to put group agents into my ontology, they do not offer a philosophical theory of group agency. It would be a stretch to attribute an account of mental states to Strauss and his followers. The present chapter fills the gap and argues for an endorsement of functionalism. The intentions, beliefs, and other propositional attitudes of group agents are functional states. Consequently, groups are agents only if they have internal states with the appropriate functional profile to realise propositional attitudes. These functional profiles, in their turn, should be the same as in the case of individual humans, although the realising states will, of course, differ in other respects. The present chapter will not specify these profiles in any detail but rely on our everyday understanding. For example, if a group has an intention to achieve a certain goal and a belief that it can be achieved in a certain way, then (other things being equal) it will try to achieve it in this way.

With functionalism in place, a third question arises: which internal states of groups realise the functional profiles of the relevant propositional attitudes? I will only provide a partial response because empirical research beyond the boundaries of philosophy is needed for a complete answer. Since functional profiles can be realised by a large variety of internal states, and groups come in diverse configurations, we have no reason to assume that there will be one identifiable type of realiser for all kinds of groups.

For example, if an egalitarian group of pre-industrial peasants and a highly developed and hierarchically structured state both have beliefs and intentions, the realisers probably vary to a large degree. In the egalitarian case, discussions, verbal agreements and the

propositional attitudes of all the members might play a more significant role than in the case of a hierarchical state. Here we should expect formal procedures and the mental states of the ruling members to play a larger role. We have no reason to expect that realisers of propositional attitudes are of one kind for all group agents.

In light of the empirical nature of this question, the present chapter cannot fully answer what kind of states in group agents realise the functional profile of the relevant propositional attitudes. It is a task for the social sciences to find the exact realising states within the various human groups. I will, however, provide a partial response drawing on the Negotiated Order approach. Its emphasis on local agreements and negotiations puts restrictions on which states are likely to realise functional states within groups, as I argue later.

But further questions remain open: the fourth asks whether group agents comply to the same theory of motivational change that Negotiated Order sociologists apply to individuals. Do group agents also respond with an exploratory phase to Problematic Situations, during which they exhibit an increased probability of preference change? In the end, this question can only be answered by empirical research since the theory describes contingent regularities in decision processes. I rely once more on the Negotiated Order sociologists and will therefore not argue further for the proposition that group agents have an increased probability of preference change in the aftermath of a Problematic Situation.

Since, as mentioned above, Negotiated Order sociologists applied their account to group agents, I follow their example. My discussion in the next chapter will reveal, however, that even if only a limited number of states in climate change negotiations fit this account of motivational change, this could already have significant consequences for how such negotiations could work. Accordingly, I only have to assume that the Negotiated Order

approach applies to a few but significant group agents for my proposal to have important consequences. Exceptions can be accommodated.

The fifth and final question follows from postulating the same theory of motivational change for individual and group agents: what is the relationship between preference change in response to Problematic Situations on the individual and on the group level? One might suggest that group agents undergo preference change because their members do so. Consider the hospital example I discussed in the previous chapter. Arguably the hospital developed a new preference for a smaller number of adolescent patients because the members of the hospital encountered Problematic Situations. The Problematic Situations and the associated preference change occurred on the group level, because it occurred on the level of individuals. But while such claims have prima facie plausibility, they have to be stated with care and qualified.

As already mentioned, groups come in a wide range of diversity and thus one should not simply presuppose that all group preferences result from an aggregation of the preferences of their members (cf. Hubener 2012: 611). For the pre-industrial egalitarian group and the small hospital, the preference change in response to a Problematic Situation might be traced back to such changes on the level of the individuals. In the hierarchical developed state, such a connection becomes more dubious. We can imagine that the state undergoes preference change because its structure is shaken up by a Problematic Situation without any of its members encountering such a situation. Only large-scale research in the social sciences will reveal whether intrinsic preference change on the group level commonly follows from intrinsic preference change on the individual level. Once more a question about group agency becomes an empirical matter and therefore beyond the boundaries of the present philosophical inquiry.

Out of the five pressing questions concerning group agency, this chapter offers a response to one and a half of them. This limitation does not result from a lack of ambition but follows from respect for empirical inquiry as the arbiter of the salient questions. Group agency and the theory of preference change will be assumed because Strauss and other Negotiated Order sociologists endorsed it. But I will argue for a functionalist construal of the mental lives of group agents, and for a restriction on which kinds of states in a group realise the functional profile of propositional attitudes.

I will support a functionalist theory of group by arguing against the main competitor: interpretivism about group agency.[65] Interpretivist theories of mind exert an immense attraction on proponents of group agency. Although their versions differ, Christian List and Philip Pettit (2011) as well as Deborah Tollefsen (2015) put forward interpretivist accounts of group agency. They endorse the principle that if one can interpret an object as an agent with predictive success, then the object is an agent.

This chapter argues that there is a class of cases which functionalism can accommodate, but interpretivism cannot. Two features characterise this class: First, distinct groups coincide, that is, numerically distinct groups share all their members at all time. Second, we have access to the inner mechanisms of the groups agents because members know what they have decided on. I construct a counterexample with these features allowing me to reject interpretivism about group agency in favour of functionalism.

First, I introduce the distinction between functionalism and interpretivism, at the heart of which lies the difference between behaviour on the outside and mechanisms on the

[65] Sometimes pragmatist sociology together with ethno-methodological approaches is called "interpretivist sociology". The sense of "interpretivist" is different and should not be confused. I always stick to the philosophical use.
Perhaps Rovane's 2014 approach to group agency can serve as a third proposal. But it has gained little traction in the literature and it is also not entirely clear whether it stands in conflict with a broad notion of functionalism.

inside. Because they limit themselves to outer behaviour, interpretivists are sometimes accused of overgenerating group agents. After having discussed the interpretivist responses to this charge, I present my own argument against interpretivism. As I show, especially challenging problems arise from examples of coinciding group agents, undermining recent defence strategies devised by interpretivists. I conclude by hinting at how functionalist accounts of group agency can succeed where interpretivists fail.

In addition, I argue for certain restrictions that a functionalist theory of group agency must fulfil to fit the Negotiated Order approach. I look at the account offered by List and Pettit from another perspective in order to provide a contrast. Their works suggest that bureaucratic procedures, such as formal votes, take centre stage in realising mental states. From the perspective of the Negotiated Order approach, we instead expect informal interactions, and in particular negotiations, to play a central role. The Negotiated Order approach presents an original account of what kind of states within groups realise the mental states needed for group agency.

With these two questions about group agency addressed and the others delegated to the appropriate social sciences, we will be able to apply the Negotiated Order approach to climate change negotiations in the next chapter.

The Functionalist Principle

Since the introduction of multiple realizability arguments, functionalism has become the standard account of propositional attitudes.[66] One can distinguish a variety of functionalisms, but they share the functionalist principle that for an entity, whether individual or group, to have states with the right functional profile is to have the corresponding propositional attitudes:

(F) An entity has propositional attitudes if and only if it has states with the appropriate functional profiles.

Ramsey sentences, that is sentences that describe the theory of propositional attitudes completely but replace mention of them with existentially quantified variables (cf. Lewis 1970), provide another way to introduce functionalism and the idea of a functionalist principle. Functionalism claims that propositional attitudes can be defined via Ramsey sentences which effectively specify their functional profiles.

For a garden-variety functionalism, such Ramsey sentences would describe the causal roles of the propositional attitudes.[67] Proponents of group agency have a natural affinity for this kind of functionalism. If all there is to having a propositional attitude is having a state with the appropriate functional profile, which specifies causal roles, then why should groups not have propositional attitudes?

Consider a coarse-grained functionalism, according to which "mental states are internal states of an agent that are caused by certain inputs to the system and cause both certain other internal states and certain behaviour outputs, where these causal dynamics will be specified by common sense." (Tollefsen 2015: 81) For example, an individual has the

[66] In line with the literature I will assume that all mental states of group agents, such as desires and beliefs, are propositional attitudes.

[67] However, a Ramsey sentence could also specify a realiser and, for example, define a certain propositional attitude as only being instantiated by a certain kind of neuron. This would be an unusual realiser-dependent functionalism, which I ignore in the following.

belief that the supermarket is on the corner because they have an internal state which makes them go to the corner when they want to buy groceries. In such a functionalism, the group mind thesis might appear plausible. A corporation could have a state that meets the functional description of a propositional attitude. For example, if it shows the behaviour of entering a market, this might be part of the functional description of having the belief that it can make profit in this market. The main difficulty lies in specifying the functional profiles correctly to attribute attitudes to the appropriate groups and individuals, but not beyond.

Functionalist approaches to group agency have a long history going at least back to Ned Block's China Brain (1978), purportedly a counterexample to functionalism, and D. H. M. Brooks' (1986) paper "Group Minds", which endorses the possibility of a city being the functional equivalent of a brain. Bryce Huebner (2014) is one of the recent proponents of a functionalist account of group agency. While he argues that the functional profile of propositional attitudes such as beliefs and desires are rather demanding, he suggests that some groups meet the requirements and become full agents. Brian Epstein (2015, 2017) has also endorsed a functionalist theory of group agency.[68]

The Interpretivist Principle

Christian List and Philip Pettit (2011) in *Group Agency*, and Deborah Tollefsen (2015) in *Groups as Agents* endorse Dennett's theory of the intentional stance in place of coarse-grained functionalism. To use the intentional stance "is to set aside non-intentional possibilities of explanation, to presuppose that the system under explanation is an agent,

[68] Epstein's version departs from the others, however. He argues that we should attribute agency on the levels of kinds of groups, rather than to individual groups. That is, we would attribute group agency to committees in general rather than look at one specific committee to discern whether it is a group agent (Epstein 2017).

and to try to ascribe representations and motivations to it that make sense of its actions" (List & Pettit 2011: 23).[69]

Consider an example adapted from List and Pettit (2011: 19-31): I look at a robot's behaviour and predict its future behaviour by ascribing propositional attitudes such as desires and beliefs. The robot moves towards some cylinders that are lying down and puts them upright. I ascribe perceptual beliefs and a desire for upright cylinders to the robot. To test my ascription, I topple a cylinder. The robot puts it upright again. Having acquired the belief that a cylinder is lying down, the robot satisfies its desire for upright cylinders by showing appropriate behaviour. My ascription results in successful predictions, which is explanatorily more powerful than a non-intentional description. Accordingly, the robot is an agent with perceptual beliefs and the desire to put cylinders upright.

In agreement with List and Pettit, Tollefsen describes interpretivism as "the view that, if we can successfully make sense of another being – understand and interpret its behaviour by using our folk psychology – it is an intentional agent" (Tollefsen 2015: 97). We interpret the behaviour of an object by ascribing propositional attitudes, and if the interpretation allows us to predict the object's behaviour, the success validates the ascriptions.

To use an example of a group, if the philosophy department's library team shows the behaviour of sorting books on its shelves, I attribute to it the desire to have its books well-ordered and the belief that the sorting behaviour helps to achieve this goal. These ascriptions allow me to predict that if I put a book from the shelf on the table, the team will put it back in the right place. The success of the predictions validates my ascriptions.

[69] See also Dennett (1987: 15; 1991).

Since ordinary people do not have access to all facts about the behaviour of entities, interpretivists assume an interpreter who is idealised in that she has access to all behavioural facts about the entity in question.

We can formulate the following interpretivist principle:[70]

(I) An entity has propositional attitudes if and only if an idealised interpreter would successfully predict the behaviour of an entity from the intentional stance by ascribing these attitudes.

The interpretivists offer their own principle for the attribution of propositional attitudes, raising the question of how it relates to the functionalist principle.

Functionalism and Interpretivism: The Outer Behaviour/Inner Mechanism Distinction

Tollefsen suggested that, in contrast to functionalism, interpretivism entails that "[p]ropositional attitudes are not internal states of a system but dispositional states of whole systems" (Tollefsen 2015: 110). Accordingly, the truth of an ascription of propositional attitudes does not depend on any facts about the internal life of the entity. In the case of the robot. its movements count, but not the calculations by its processor unit. In the case of the library team, moving the books onto the shelf counts, but not the internal deliberation about where to put them.

Functionalist accounts typically do not care whether agents are made of carbon or silicon, but they often demand that internal states fulfil certain functional roles to realise

[70] One might emphasise in addition that the intentional ascription has to provide more predictions than other available approaches, but I leave this out of the principle to keep it simple.

propositional attitudes. Tollefsen denies such internal realisers any role for defining propositional attitudes.

The distinction becomes slightly muddled in List and Pettit's work, who not only endorse Dennett's intentional stance theory like Tollefsen, but also what they consider a functionalist analysis of "intentional states – beliefs and desires – in terms of the roles they play in directing the agent and guiding action" (List & Pettit 2011: 171). But since their endorsement of Dennett's intentional stance is beyond doubt (see List & Pettit 2011: 11, 13, 23), we should understand List and Pettit as using an unusually broad notion of functionalism, which does not commit to the existence of internal states and is therefore compatible with interpretivism.[71]

In the following I use "functionalism" for non-interpretivist versions of functionalism, that is, versions of functionalism according to which the Ramsey sentences of propositional attitudes involve internal states. Thus, the separation of the two theories relies on a distinction between outer behaviour and inner mechanism.

For individuals giving an approximate criterion for the outer behaviour/inner mechanism distinction proves easy:[72] Everything that stays within the skull is part of the inner mechanism rather than the outer behaviour. We might make some exceptions, but as a general heuristic the criterion will do.

The intracranial criterion, however, proves unsuitable for drawing the distinction with regard to groups. Groups do not have a skull of their own. Neither List and Pettit, nor

[71] Tollefsen reads List and Pettit as suggesting "that the formation of group judgements […] somehow realizes group beliefs" (Tollefsen 2015: 81). According to Tollefsen, List and Pettit endorse a coarse-grained functionalism in which the functional profile includes internal mechanisms. I consider this a misreading, although one that can be made productive as I show later.

[72] This matter is complicated by the debate on the extended mind, which is also discussed by Tollefsen as supporting group agency (2006, 2015). In the following, I will leave this complication aside.

Tollefsen, offer an explicit criterion suited for groups, leaving it to their readers to judge where inner mechanism ends and outer behaviour starts. But while explicit limitations are lacking, the dialectic of the debate imposes restrictions on how to construe the outer behaviour/inner mechanism distinction.

For interpretivism about group agency to be interesting, enough events need to fall into the inner mechanism category. An overextended category of outer behaviour threatens to render the difference to functionalism negligible. Interpretivism about group agency would not be of interest if it used all information about documents and discussion within the group as the basis for attributing propositional attitudes.

In the case of individuals, any neural behaviour falls into the inner-mechanism rather than the outer-behaviour box. But then for groups, does everything that is individual behaviour fall into the inner-mechanism box too? One should not push the analogy between neurons and individuals too far. While Tollefsen (2015: 106-107) suggests that an interpretivist could ignore some individual behaviour, for example they could ignore the actions of individual managers to predict the Ford Motor Company's response to an increase in gas prices, ruling out all individual behaviour goes too far.

If we ignore the behaviour of all members of the library team, no group behaviour would remain. The library team showed the behaviour of sorting books, but so did the relevant individuals. At least for groups exhaustively constituted by individuals we want to allow that an event can be both outer behaviour of the group and individual behaviour. However, this individual behaviour should not be internal to the group, but rather relate it to the outside. In the book sorting scenario, the individual members engage with the books, which are external to the library team. A deliberation about book sorting, however, would remain internal, because the members only engage with one another.

Let me then propose the following criterion: A behaviour of a constituent[73] of a group is an instance of an outer behaviour of the group if and only if the behaviour engages with an entity external to the group. Only those outer-behaviour events of constituents of the group are outer-behaviour events of the whole group, which involve non-constituents as well.

Without limiting which individual behaviour counts as outer-group behaviour in such a way, the interpretivist approach to group agency is not interestingly different from its functionalist rival. After all, the functionalist is likely, at least in some cases, to identify the internal states of group agents realising propositional attitudes with states of individuals and their behaviour. The constituents of these states can then no longer serve as the basis for applying the intentional stance on pain of rendering interpretivism and functionalism effectively equivalent.

Of course, many difficulties remain for applying the distinction. My proposed criterion suggests that a conversation between group members would be outer behaviour if someone from outside the group happened to participate in it, but not if the same contributions were made by a group member. Furthermore, my criterion leaves open what counts as an engagement or involvement of non-constituents. Moving air or radiating heat are not enough, otherwise internal deliberations of groups would also count as outer behaviour. After all, our neuronal activities also create external traces and interpretivism does not consider them behaviour. But glossing over such difficulties is only charitable towards the interpretivist. If it turned out that there is no principled criterion for distinguishing outer behaviour from inner mechanisms, then this would just settle the issue in favour of functionalism.

[73] I use the notion of a constituent rather than a member to allow that more than individuals might constitute a group and contribute to the behaviour of the group.

The Overgeneration Worry

While only taking predictive success about outer behaviour into account distinguishes interpretivism from functionalism, it raises the concern that interpretivism hopelessly over-attributes group agency. Will we not always have some predictive success using the intentional stance towards groups? Might I not predict some outer behaviour of the global human population using intentional vocabulary? Although I can roughly predict the behaviour of humanity by ascribing it the intention to increase the global temperature, it does not form a group agent, certainly not one intending global warming.

Dismissing interpretivism proves more difficult, however, because interpretivists demand sophisticated behaviour for agency. List and Pettit (2011: 24-25) introduce strict rationality criteria, for example that attitudes must track facts, that attitudes be internally coherent, and that actions must follow attitudes. It is doubtful whether the behaviour of humanity allows such rational interpretation. For most intentions we might ascribe to humanity, we will find considerable violations of rationality. While we have some predictive success attributing the intention to increase the global temperature, some of humanity's behaviour, such as a reduction of coal plants, conflicts with it.

Tollefsen (2015: 102, 108) reminds us that Davidson suggested linguistic intelligibility as an interpretivist requirement. An agent must engage in linguistic behaviour which we can interpret. Since we cannot attribute any rational linguistic output to humanity, it is no agent. In effect, Tollefsen further narrows what counts as a successful predictive interpretation from the intentional stance: It has to include successful predictions which are either based on or concern linguistic behaviour.

While I am willing to grant to interpretivists that they have found ways to avoid the *overgeneration* of group agents, they cannot evade the problems resulting from *coinciding* group agents.

The Counterexample

The interpretivist principle implies that any difference between the mental lives of two group agents is to be justified with reference to the outer behaviour of the groups. This consequence of the interpretivist principle creates a problem for certain cases of coinciding group agents.

Consider two coinciding groups, which are completely constituted by human individuals. Perhaps no such individualist constitution holds for Tollefsen's example of the Ford Motor Company, but imagine that a philosophy department assigns each of its members randomly to committees, where such committees are exhaustively constituted by their individual members. The department throws all the names of the faculty in a big urn and picks five names from it for a committee. It then throws the five names back in and draws again for the next committee. By pure chance, the teaching committee might end up with the same members as the public engagement committee.

In such a case, each event that is an outer behaviour of the one group is also an outer behaviour of the other group since the groups are only constituted by individuals. They have no other constituents which could realise the behaviour of the group. "Behaviour" must be understood here in a thin sense, preceding any attribution of mentality, since only such behaviour can serve as the basis for attributing mentality using the intentional stance. It follows that when the teaching committee shows the behaviour of sending out

a resolution about pedagogical methods, the public engagement committee shows the same behaviour.[74]

Given that the two groups show the same behaviour, the question arises as to which of the two committees had the intention to distribute the resolution. Assume that it was the teaching committee, not the public engagement committee, which intended to send out the email. How can an interpretivist justify this difference in propositional attitudes between the groups, given their behavioural coincidence?

While it becomes difficult for interpretivists to pry the mental lives of two coinciding group agents apart, they can still respond to this example. The behaviour of the two groups might be indistinguishable, but the features of the behaviour clearly indicate which committee is acting in sending out the resolution. The appropriate email might be signed by the teaching committee. Tollefsen suggested linguistic interpretability as a condition for attributing propositional attitudes, and linguistic behaviour can help the interpretivists out again.

A variation of the example undermines this response. Assume that the teaching committee forms an intention to play a prank on the department. The group decides to hide alarm clocks, which will disrupt lectures. The deliberation about the prank is an inner-mechanism event which results in an outer behaviour of hiding clocks. However, the group never shows any behaviour revealing that it was the teaching committee rather than the public engagement committee which decided to play the prank. The deliberative behaviour of the group members does not count as group behaviour since, as I discussed

[74] One option I am not considering here and in the following is that groups might exist intermittently, so that the teaching committee only exists whenever the public engagement committee doesn't. While this would work against my examples, they could easily be adapted to avoid this issue.

above, a behaviour of a constituent of a group is an instance of an outer behaviour of the group if and only if the behaviour engages with an entity external to the group.

The inner mechanism of the group deliberation would give us a clear answer to which group intended the prank: the teaching committee. The members decided to play the prank as the teaching committee during one of its meetings, not during a meeting of the public engagement committee. The interpretivist, however, cannot let this inner mechanism make the difference at the risk of becoming indistinguishable from the functionalist. At the same time, the outer behaviour, the hiding of alarm clocks, does not allow a clear attribution of the intention to the one committee rather than the other.

At this point, interpretivists feel some pressure but they can still go dispositionalist. In addition to actual behaviour, behavioural dispositions might count in the attribution of propositional attitudes. The two committees have the disposition to show behaviour clarifying which committee played the prank under appropriate circumstances. Prompted by an inquiry, perhaps enforced by threats of being fired, the individual members would say that the teaching committee hid the alarm clocks.

Tollefsen's quote that "[p]ropositional attitudes are not internal states of a system but dispositional states of whole systems" (Tollefsen 2015: 110) suggests such a dispositionalist interpretivism. However, the dispositionalist move risks making interpretivism uninteresting as a position if taken too far.

As I emphasised, interpretivism is substantially different from functionalism because it avoids giving the inner mechanisms any role in ascribing propositional attitudes, and rather focusses on the prediction of behaviour. Accordingly, an interpretivist cannot just ascribe behavioural dispositions based exclusively on the inner mechanisms. At least in one plausible scenario, the entity in question must *show* the behaviour. Since interpretivism takes the attribution of propositional attitudes to be all about prediction

of behavioural patterns, pointing to an unrealisable disposition, for example a finkish disposition, is an illegitimate move within the framework. Only by limiting themselves in such a way can interpretivists turn dispositionalist without becoming uninteresting.

For the previous example this limitation does not matter. The teaching committee has a plausibly realisable disposition to reveal its authorship of the prank. The nature of group agency, however, allows us to construct a revised counterexample of coinciding groups.[75]

Assume that for some arcane reason of university bureaucracy, both the teaching committee and the public engagement committee have the capacity to dissolve themselves and each other by simply intending to do so. At any point in time, each group can end its own existence or that of the other group by forming the appropriate intentions. Let us furthermore assume that both groups have a joint session at the end of which both groups end their tenure. The two intentions are instituted by a committee member stating: "It is hereby decided that the teaching committee intends to end its tenure and the public engagement committee intends to end its tenure."

Both groups came to an end, but how could interpretivism settle which group intended to dissolve which? Did each group intend to dissolve itself, or the other, or perhaps one group intended to dissolve both? The interpretivist cannot tell. The group life ceases immediately with the act of forming the intention, otherwise resulting behavioural disposition might be realised. As sketched in my counterexample, neither committee could ever realise a behavioural disposition following from its intention.

One might try to solve the problem by pointing to behavioural dispositions prior to forming the intention.[76] For example, the groups might have a disposition to clarify which intentions they are about to form. Just interrupt the speaker after "It is hereby decided…"

[75] The following counterexample has been improved with considerable help from Luca Barlassina.
[76] I thank Yonatan Shemmer for raising this problem.

and before the end of their sentence and ask about what will be decided. However, such cases can also be ruled out with a simple tweak, namely the introduction of some randomness. Let the group member say: "It is hereby decided that if the coin comes up heads, the teaching committee intends to end its tenure and the public engagement committee intends to end its tenure, and if it then comes up tails, the intentions will be the other way round." Now neither dispositions prior to the formation of the intentions, nor any later dispositions, can reveal the content of the intentions.[77] At best the combination of a prior disposition plus facts about the coin can do so. For all practical purposes, however, whether the coin comes up heads or tails does not constitute outer behaviour of the groups.

I do not deny that the group members exhibit a different behaviour because of the groups' intentions. Asked which group formed which intention, they can clarify that each group intended to end itself and not the other. But my neurons also exhibit behaviour and dispositions when I form an intention, and interpretivism is committed to ignoring them. Only the realisable outer behaviour of the entity in question counts, and the groups cannot show such behaviour after the end of their existence. Interpretivism does not have the resources to justify the correct attribution of different intentions to the two group agents.

To summarise my counterexample, we have here a difference in the mental lives of the two committees, each intended to end itself rather than the other, although each could have intended to dissolve the other. The interpretivist faces a puzzle: Which group formed what intention? The group behaviour and realisable dispositions do not allow us to answer the question.

[77] I used the coin example for simplicity. If we assume that some randomness occurs in the decisions of individuals, it would suffice to let them make a decision.

The two special features of group agency pose a particularly difficult challenge to the interpretivist: Since the groups coincide and are exhaustively constituted by human individuals, they show no behavioural difference. Nonetheless, the interpretivist will find it hard to deny the mental difference between the two groups because we have access to the mental lives of the groups. We can simply ask the members about the meeting.

Four Interpretivist Responses

An interpretivist might respond in four ways to my challenge. First, they might be tempted to deny that we have two groups and try to collapse them into one group agent. They would attribute the intention to end the committee tenures to the one overarching group.

But collapsing the groups goes against the growing consensus in the group ontology debate (see Gilbert 1992: 220-221, Uzquiano 2004, Sheehy 2006, Ritchie 2013 and 2015, Thomasson 2016): Groups which share all their members can remain numerically distinct. The teaching committee has duties that the public engagement committee does not have. The groups differ in their properties. Leibniz's law dictates that they cannot be one group. We have to attribute the different intentions to separate groups.

Second, interpretivists might accept that the groups are numerically distinct but suggest that both groups, the public engagement committee as well as the teaching committee, formed intentions to end both groups. Here I rely on our intuitions about the case. The individuals discuss the two intentions and know which group formed which intention since they debated it.

Given these details about the meeting, that is, the inner mechanisms of the groups sustaining the outer behaviour, it appears wrong to attribute the same propositional

attitudes to the two distinct groups. But the inner mechanisms are not allowed to play a role in the attribution of propositional attitudes according to interpretivism.

Third, interpretivists could attempt to deny that the groups formed any intentions at all, because they show neither the required outer behaviour nor realisable behavioural dispositions. This solution might have some plausibility for individuals, where interpretivists might assert that if an individual shows no behaviour and not even behavioural dispositions indicating an intention, we have no reason to attribute one.

There are, however, two reasons why this response fails for the counterexample. The first reason is that the groups ended. If there were no intentions to this effect, we have no explanation regarding why the committees went out of existence. The second reason returns to the special feature of group agency: Group members know what they decided on and in particular which committee intended what. Their knowledge derives from their involvement with the inner mechanisms of the group. In the case of group agents, especially those only constituted by individuals, we have a kind of access to the inner life of the agent, which we do not have for individual agents. This peculiarity of group agency allows us to pre-empt the third interpretivist response.

As a fourth and last refuge, interpretivists could argue that I have drawn the line between inner mechanism and outer behaviour wrongly. If the deliberation between the teaching committee members, their discussion about which group intended what, were to count as outer behaviour of the committees, then we could account for the difference in the mental lives. The behaviour of the members deliberating for the committees would allow the interpreter to make the correct attribution.

But this retreat renders interpretivism about group agency uninteresting. Interpretivism about individuals is a substantial thesis because it stops, roughly, at the skull. As discussed, interpretivism about group agents is only interesting if it endorses an analogous limit.

Tollefsen (2015: 106-107) herself recognises this in her discussion of the Ford Motor Company: the interpretivist position is interesting because it allows us to ignore the behaviour of individuals, the discussion between the president of the company and other members, and exclusively considers the outer behaviour when we attribute to the company the intention to raise prices. If the deliberations between the members also go into the outer-behaviour box, then little difference to functionalist accounts of group agency remains. Similarly, an interpretivist who responds to worries about misattributing intentions to an individual cannot just solve the issue by categorising neuronal processes as behaviour of the individual.

In conclusion, my counterexample of the coinciding groups establishes that interpretivists cannot offer an interesting account of group agency and correctly attribute different mental lives to coinciding group agents in all cases.

The Underlying Flaw and Another Failed Response

My counterexample exposes an underlying flaw of interpretivism: Interpretivism wrongly limits the basis of interpretation to outer behaviour. Other counterexamples against interpretivist theories of individual agency also rely on this flaw. Lycan (1987: 5) proposed the tinfoil man against behaviourist theories of mind,[78] Peacocke (1983) brought in Martian marionettes, and Block (1981) has his Blockheads. These are all variations on a theme. I focus on the Martian marionettes, since Peacocke explicitly directs the example against Dennett's theory.

Peacocke invites us to imagine a human body without a brain; instead a radio transmitter controls the nerves. A computer on Mars calculates the behaviour and manages to make

[78] Bryce Huebner (2014: 90) has pointed out that the tinfoil man example speaks against attributing agency to groups based only on the intentional stance.

the body, its marionette, behave like an ordinary human. As Peacocke observes, the marionette "is voluminously and reliably predictable via the intentional strategy, as voluminously and reliably as for any normal human being" (Peacocke 1983: 205). Interpretivism gives the wrong conclusion for the Martian marionette. It finds an agent, where we only see a puppet.

For Martian marionettes and coinciding groups, looking only at the outer behaviour gives the wrong answer. In both cases, behaviour does not suffice for our attribution of propositional attitudes. We have to look under the skin. Thus, one might hope that the interpretivist response to Martian marionettes also solves the problem of coinciding group agents.

Bruno Mölder, whom Tollefsen (2015: 97) quotes as a recent defender of interpretivism, has responded to the marionette-type counterexamples. He admits that looking at predictive success is insufficient, but suggests that folk psychology imposes restrictions on the possible objects of the intentional stance. According to Mölder "[i]t is part of folk psychology that people have beliefs whereas tables and lecterns do not" (Mölder 2010: 193). The Martian marionette does not fall into the range of objects covered by folk psychology, because it does not take objects with empty heads seriously. If folk psychology only covers a limited range of objects, then it limits our interpretations and rules out Martian marionettes.

Leaving aside whether Mölder's response to the Martian marionettes succeeds, it fails against my counterexample. Mölder objects to the far-fetched sci-fi character of the counterexamples. But in the case of the teaching and the public engagement committees we do not have an empty head and radio controls. We have two groups, which look like agents with different mental lives. Folk psychology rules out neither group as an odd fringe case. Excluding these two committees as weird from the scope of interpretivism

would be tantamount to admitting that the interpretivist theories of group agency offered by Tollefsen and List and Pettit are broken.

While my counterexample hinges on the same underlying flaw as the Martian marionette example, interpretivist theories of group agency cannot evade it using Mölder's strategy. The two committees differ in their mental lives and the members know it from the inside even though it never shows in behaviour or behavioural dispositions of the groups. Although interpretivism has notable followers in the debate on group agency, it fails particularly badly in this area.

The Functionalist Solution

If interpretivism cannot distinguish between the mental life of the two committees, how do we pull it off? We know that the teaching committee intended to dissolve itself, because we know that its members discussed this at the meeting. By looking more closely at the interaction of members, we can pry apart the mental lives of the two groups. A functionalism which considers *how* the members realise the group's propositional attitudes provides the correct answer.

Tollefsen rejects such a functionalism claiming that "[p]ropositional attitudes are not internal states of a system but dispositional states of whole systems" (Tollefsen 2015: 110). For the case of the two committees, the internal states of how the members realise the groups' mental lives make all the difference, while the outer behaviour, realised and dispositional, of the two committees remains indistinguishable.

List and Pettit do not fare any better since they, too, state that "the performance itself should dictate the representations and motivations we ascribe to the agent" (List & Pettit 2011: 28). For the case of coinciding group agents, this principle fails since the

behavioural performances of the groups remain indistinguishable. Only a functionalism that looks beyond the performance of the system and considers how internal mechanisms realise it can solve the problem.

Functionalist proponents of group agency have yet to explain why the deliberation between group members matters for the realisation of the teaching committee's intention. They must specify the Ramsey sentences in such a way that the intentions are attributed to the appropriate committee. For example, the Ramsey sentence for an intention might demand that the realiser stands in a certain causal connection to a realiser of the agent's self-representation. This link to a self-representation of the group would allow the functionalist to argue why the teaching committee forms one intention and the public engagement committee the other: there are two self-representations for each group and each of them only stands in an appropriate relation to one of the intention-realisers. In the case of the coinciding group agents, the connection might be realised through the individual members and how they conceive of their deliberation. In the case of individual agents, the realiser would presumably not involve other agents.

Of course, the functionalist would have to justify such a requirement independently of the presented problem case so as not to appear terribly ad hoc. However, the mere possibility of such a response renders functionalism the more attractive theory of group agency. While functionalist proponents of group agency face the challenge of providing the correct Ramsey sentences, interpretivists close the door to a solution.

If interpretivists want to persist in basing the attribution of propositional attitudes on group behaviour, they must find a better response to the problem of coinciding group agents. For now, the functionalist proponents of group agency remain vindicated. I therefore conclude that sociologists who adopt the Negotiated Order approach should

endorse a functionalist theory of mind, for which one can accept Tollefsen's characterisation:

"According to coarse-grained functionalism […], mental states are internal states of an agent that are caused by certain inputs to the system and cause both certain other internal states and certain behaviour outputs, where these causal dynamics will be specified by common sense." (Tollefsen 2015: 81)

In my proposal, we identify internal states of groups as their mental states based not only on the behaviour of the group, but also on facts about the internal mechanisms of the group. In particular, facts about how the individuals conceive of the group's behaviour figure in the realising states. This way, my broad functionalism can avoid the problems with coinciding group agents.

The kind of functionalism I have hinted at also fits exceptionally well with the Negotiated Order approach. Like symbolic interactionism, in the Negotiated Order approach, the way agents interpret their actions matters greatly. In chapter three I discussed the importance of meaning in interactions for the Negotiated Order approach. As I just argued, for the example of the two committees the interpretations of the individuals make the difference. My functionalism and the Negotiated Order approach both give a role to individuals interpreting group behaviour.[79]

Intentional stance theory would not have exhibited the same affinity to a theory of action endorsed by the Negotiated Order approach. It would have reduced the interpretations of the group members to a secondary phenomenon. The outward behaviour of the group would have determined the assignment of agency. My proposal combined functionalism and the Negotiated Order approach in a congenial unified theory of group agency. But

[79] This does not mean, however, that the attitudes of the group follow from an aggregation of the individual attitudes.

the Negotiated Order approach allows us to say even more about the functional realisers for group agency.

The Negotiated Order Approach and the Functional Realisers of Mental States

Since functionalism remains typically neutral concerning what realises mental states, it offers itself as a theory of group agency. A functionalist approach can allow that silicon chips in a robot to realise the same mental states as the neural network in a human head. As long as the functional profile of the mental state is realised, we have a mental state of the same type.

Endorsing functionalism, defenders of group agency only have to define what realises the functional role of mental states in the group. Neural networks in our heads presumably have the functional profile needed for having mental states; the proponents of group agency have to find an analogue realiser in groups. At this point we only know that, at least in some cases, how individuals conceive of the group behaviour makes a difference for the realisation of mental states, because otherwise we run into problems with coinciding group agents. That leaves multiple proposals on the table.

One salient option is to identify the functional realisers of mental states with bureaucratic rules and official decisions within the groups. The Weberian account of organisation provides the basis for such an identification. Although they endorse interpretivism, List and Pettit's work on group agency also suggests such an identification of the functional realisers of mental states with bureaucratic elements of organisations.

Tollefsen reads List and Pettit as suggesting "that the formation of group judgements […] somehow realizes group beliefs" (Tollefsen 2015: 81). And while List and Pettit clearly

endorse intentional stance theory, their influential book *Group Agency* focusses on various voting procedures for judgement aggregation. In the following, I ignore List and Pettit's endorsement of intentional stance theory and reinterpret them as endorsing a broader functionalism. In this interpretation, they could accept that the facts about how the individuals interpret the bureaucratic elements belong to the realising mental states. We would find a broad functionalism combined with a focus on bureaucratic elements. I present their approach to reveal the tension between such a bureaucracy-focussed picture of group agents and the one suggested by the Negotiated Order approach.

List and Pettit's work on the internal functioning of groups revolves around the aggregation of individual propositional attitudes to that of a group agent. They discuss at great length which aggregation functions might allow group agency. An appropriate aggregation function takes the attitudes of the group members as input and returns the attitudes of the group as output in such a manner that the group functions as an agent. Most of their discussion concerns one desideratum for the output: the resulting propositional attitudes must be consistent enough to guide action.

This desideratum follows from functionalism and the identification of the aggregation result with the attitudes of the groups. Unless the result shows sufficient consistency to guide agents, it cannot realise mental states. The functional realisers of mental states must meet the consistency requirements that come with the functional profile of the mental states they are realising. Only because the aggregation result is supposed to realise the mental state and determine its content, does the consistency requirement make sense. It proves a challenging requirement.

Assume that a group of three people has to take an epistemic attitude towards three propositions.[80] The third proposition is the conjunction of the first two propositions. All three members vote on the truth of the propositions. The first group member votes for the first two propositions and the conjunction as a belief. The second and the third member each endorse a different one of the first two propositions and accordingly reject the conjunction. Using majoritarian aggregation to establish which propositions the group believes, the group might end up with a result where it endorses the first two propositions as beliefs but rejects the conjunction. The result can be represented in a table much employed by List and Pettit:

Members	Proposition 1: p?	Proposition 2: q?	Proposition 3: (p&q)?
Member 1	p	q	(p&q)
Member 2	p	not-q	not-(p&q)
Member 3	not-p	q	not-(p&q)
Majority	p	q	not-(p&q)

The resulting beliefs hold the following propositions true: p, q, and not-(p&q). As we can see, the group ends up with inconsistent beliefs while all the group members have consistent sets of propositional attitudes.

This situation is an instance of what List and Pettit call "the discursive dilemma". It shows that majoritarian aggregation does not meet the first desideratum since the propositional attitudes resulting from the majoritarian aggregation function are inconsistent and therefore fail to guide action. Even worse, the discursive dilemma not only arises for a

[80] The same dilemma arises for group intention. I only focus on the epistemic case to keep the discussion simple.

majoritarian aggregation, but for a wide range of aggregation functions (cf. List & Pettit 2011: 47-50). List and Pettit prove that no aggregation function can have the following four features (see List & Pettit 2011: 49):

- **Universal domain**. The aggregation function admits any possible profile of individual attitudes towards the propositions on the agenda as input, as long as the individual attitudes are consistent and complete.
- **Collective rationality**. The aggregation function produces consistent and complete group attitudes towards the propositions on the agenda.
- **Anonymity**. All individuals' attitudes have equal weight in determining the group attitudes. Formally, the aggregation function is invariant under permutations of any given profile of individual attitudes.
- **Systematicity**. The group attitude to each proposition depends only on the attitudes the individuals have towards it, not on their attitudes towards other propositions, and the pattern of dependence between individual and collective attitudes is the same for all propositions.

For an aggregation function to meet the desideratum, at least one of these four demands must be relaxed. List and Pettit go through them in turn (see List & Pettit 2011: 51-58). For my purpose of illustrating the bureaucracy bent of List and Pettit's work, it suffices to look at the main solution they tend towards.

List and Pettit's preferred solution is to relax systematicity and includes giving up the requirement that the pattern of dependence between individual and collective attitudes has to be the same for all propositions. An example of such an aggregation function is the premise-driven procedure, which "generates a group attitude towards each premise by taking a majority vote on that premise and then derives its attitudes on the conclusions

from its majority attitudes on the premises" (List & Pettit 2011: 56). To use the table from before, the results change as follows:

Members	Proposition 1: p?	Proposition 2: q?	Proposition 3: (p&q)?
Member 1	p	q	(p&q)
Member 2	p	not-q	not-(p&q)
Member 3	not-p	q	not-(p&q)
Result	P (majority vote)	q (majority vote)	p&q (derived)

As shown in the table, this procedure ensures a consistent output. The premise-driven procedure illustrates a set of aggregation functions that meet the first desideratum by giving up systematicity. It serves as a proof of concept that aggregation functions can meet the desideratum for group agency. They take the attitudes of the group members as input and return the attitudes that are consistent enough to guide action. Such aggregation functions are not essential for structuring group agents, however.

Considering how much effort List and Pettit spend on specifying which aggregation functions can establish group agency, their admission that group agents do not have to be structured around such functions might come as a surprise. List and Pettit distinguish between aggregation functions and the organisational structure of a group agent, where organisational structures are "the rules and procedures the group uses to implement, and subsequently to enact, such a[n aggregation] function" (List & Pettit 2011: 60).

Organisational structures can be based directly on an aggregation function but they do not have to be. List and Pettit draw a distinction between such functionally explicit

organisational structures and inexplicit ones. An organisational structure is functionally explicit, "if the group explicitly uses a given aggregation function […], applies it mechanically to the attitudes of its members, and then enacts the resulting group attitudes in an equally mechanical way" (List & Pettit 2011: 60). By contrast, functionally inexplicit structures do not commit to one aggregation function, but rather involve "a heuristic for determining, from proposition to proposition, the way for the group to go on" (List & Pettit 2011: 61).

List and Pettit repeatedly use a straw-voting procedure as an example for inexplicit organisational structures. In a straw vote, the group roughly follows these steps (for a more extensive version, see List and Pettit 2011: 62):

- Consider the relevant propositions one by one.
- Take a majority vote on each proposition.
    - If the attitude formed is consistent with attitudes already formed, it becomes the group attitude.
    - If the attitude formed is inconsistent with the attitudes already formed, have a vote on possible revisions until a consistent outcome is reached.
- Assign suitable members to enact the resulting group attitudes.

This straw-voting procedure ensures that the group agent ends up with consistent propositional attitudes for all relevant propositions if it forms a profile at all. List and Pettit do not demand that group agents explicitly structure themselves around an aggregation function, but their functionally inexplicit organisational structures must fulfil the same demands of consistency.

The reasons for this limitation follow from List and Pettit's identification of the states resulting from these organisational structures with the mental states of the group.

Realisers for functional states have to meet consistency requirements because otherwise they would fail to have the appropriate functional profile for states, such as beliefs and intentions. But if the outcome of the vote did not fix the content of the mental state, then there would be no justification for a consistency requirement on the voting procedure.

List and Pettit identify mental states with the states resulting from organisational structures, which they illustrate with examples of bureaucratic procedures. In their examples, they identify the content of these mental states with the results of bureaucratic voting procedures. These identifications are compatible with the endorsement of a functionalism which resolves the problem of coinciding groups. List and Pettit's discussion suggests a focus on the bureaucratic structure even on the assumption that my previous argument for broad functionalism succeeds.

Having familiarised ourselves with List and Pettit's position, a stark contrast to the Negotiated Order picture becomes apparent. While the Negotiated Order approach allows bureaucratic rules a limited role, List and Pettit base their whole account on strict aggregation functions and voting procedures. Even their favourite example for a functionally inexplicit structure has a formal character. The Negotiated Order approach instead emphasises the need for informal negotiations and on-the-spot working agreements, which the group members revise often and without much reference to official rules or votes.

The difference between the two accounts matters for applying functionalist accounts of mind to groups. List and Pettit's discussion suggests that the functional realiser is a formal aggregation or voting procedure that results in an explicit endorsement of a belief. By contrast, the Negotiated Order approach only accepts such a realisation of group beliefs for marginal cases. Strauss and his team argued that the rules governing the behaviour of the professionals in the hospital "are far from extensive, or clearly stated, or clearly

binding" (Strauss et al. 1963: 151). Facts about these rules do not function as the action guiding states that mental states should.

While List and Pettit might perhaps grant that procedures of establishing group beliefs are more informal, the problem cuts deeper. The whole picture of official results guiding the whole behaviour of the group is flawed. Rules and official results can play a role, but as Strauss and his team found, "the area of action covered directly by clearly enunciated rules is really very small" (Strauss et al. 1963: 153). For the regular functioning, local negotiations and agreements matter more.

The Negotiated Order account suggests that we will not find the functional realisers of propositional attitudes in bureaucratic rules, formal voting procedures, and official declarations. The functional realisers are rather aggregate states, which can include states concerning such rules, procedures, and declarations, but more importantly on-the-spot negotiations, local agreements, and unspoken understandings. Reading the propositional attitudes from the official voting results proves impossible in this picture. The official record, voted upon in a straw-vote procedure, might say one thing, while what guides the group are on-the-spot negotiations with another result.

Accordingly, we do not have to expect that the group's official declarations and voting results meet the kind of consistency that List and Pettit expect. Whatever rationality restriction we impose on the content of mental states, they are not therefore restrictions on the formal procedures, but rather constrain the more complex realisers, which include negotiations, local agreements, and unspoken understandings. Determining the exact realisers remains a task for the social sciences.

Having looked inside group agents, we now turn our attention to their interactions. Negotiated Order sociologists suggest that group agents, too, encounter Problematic Situations, open up to intrinsic preference change, signal in their exploratory phase, and

negotiate with one another. The Negotiated Order approach's theory of action applies to group agents and the next chapter explores the consequences for the special case of climate negotiations.

Chapter Six: The Example of Climate Change Negotiations

Having combined the theories of Negotiated Order and group agency, I have put forward a theory of negotiating group agents. At the beginning of my thesis, I motivated developing such a theory by pointing to the large-scale problems humanity faces and which it might hope group agents will address through negotiations. In this chapter, I apply the theory of negotiating group agents to a particular large-scale problem: climate change.

My promise was that the Negotiated Order approach to climate negotiations has a contribution that goes beyond what standard rational choice models offer. If we take the perspective of Negotiated Orders, we find different opportunities for influencing climate change negotiations so as to achieve a positive agreement.

The first half of this chapter discusses standard rational choice models of climate change. Since the literature provides an abundance of such models, I limit myself to a plausible and relevant subset drawing on the work by Stephen Gardiner. After looking at these models, I turn to the difference that the Negotiated Order approach makes. I introduce the mechanism of preference change and signalling that I discussed in chapters two and three and apply them to the case of climate negotiations between states. Based on these discussions I provide some general proposals for how to achieve an ambitious climate agreement.

The result shows that although the Negotiated Order approach has some shortcomings of its own, such as requiring further formalisation and empirical input, it also has contributions to make beyond what rational choice models can offer.

Standard Rational Choice Models of Climate Change

A great variety of standard rational choice models try to represent responses to climate change. They come with different assumptions, different degrees of complexity, and take different entities as their basic agents. I limit my discussion to models which assume that states are group agents negotiating with one another in climate negotiations. All models describing reactions to climate change on the level of individuals remain outside the scope of my work.

Despite this limitation, we still face an impractically high number of models. With eyes on climate negotiations, DeCanio and Fremstad (2013) showed that just focussing on plausible 2x2 games in which we have two agents which can either pollute or abate, we end up with 25 possible games. The authors group these games into six groups, but that hardly solves the problem of complexity since the assumption that we only have two agents and two possible actions is dubious. Climate change negotiations concern a large number of agents even if we only take states into consideration. There are also more than two actions the states can take since emissions come in degrees. Without these assumptions, however, the number of models increases beyond what we can handle comprehensively.

Since I cannot deal with all these models, I instead pick especially conspicuous ones for illustration and informally note various complications. Taking these complications into account one can produce a myriad of available models.[81] My discussion closely follows Stephen Gardiner's influential work on climate change negotiations. In his book, *A Perfect Moral Storm,* Gardiner presents two scenarios for climate change negotiations, one

[81] One kind of model, which has received only little attention in philosophical debates, are bargaining games trying to model the learning dynamics in climate change negotiations (see Smead et al. 2014). In effect, these models introduce a further problem: even if there are satisfying equilibria available, the players are not guaranteed to achieve them. However, these models are also based upon the assumption of fixed preferences.

optimistic, one pessimistic. The two scenarios resemble two well-known 2x2 games: the battle of the sexes and the prisoner's dilemma.

Following Gardiner, I look first at the optimistic case and the reasons why we should assume the pessimistic scenario. Afterwards I informally add the complications to the pessimistic scenario. But Gardiner only provides a descriptive model of climate negotiations that leaves out important advice that rational choice theory offers for structuring such negotiations. To fill this gap, I look at a model proposed by Boadway et al. that suggests a way to achieve better negotiation results.

Optimistic Model: The Battle of the Sexes

Gardiner bases his optimistic scenario on the battle of the sexes. The standard examples for battles of the sexes describe two agents who prefer different activities but most of all prefer to do something together. Bob and Jess want to start a reading group together. Bob prefers reading Hegel over reading Kant and Jess prefers Kant over Hegel. But most of all, the two want to read a text together. We can represent the scenario in the following matrix:[82]

Battle of the Sexes		Jess	
		Hegel	Kant
Bob	Hegel	3 (Bob), 2 (Jess)	1, 1
	Kant	0, 0	2, 3

[82] The model is not symmetric because both Bob and Jess derive some utility from reading their favourite author, even if they do not do it together.

Bob and Jess have a mild conflict of interest, insofar as they do not agree trivially on what to read, but it is in both their interests to settle on one author. The outlook for an agreement is promising. The solutions come easily since, as Gardiner notes, "the collective action problem may be resolved without the need for any change in payoffs or motivation on the part of the players" (Gardiner 2011: 88). Jess and Bob both have an incentive to acquiesce if the other remains stubborn.

To apply this game to climate change negotiations we must extend it to multiple players, since multiple states negotiate their emission targets. For this purpose, Gardiner presents the example of a group of people who want to organise a game of rugby (see Gardiner 2011: 89). They must form two teams for the game. All participants prefer to watch rather than to play, but most of all they want the rugby game to take place. We now have multiple players, but also a structure that resembles the original battle of the sexes. Importantly, we can remain optimists about cooperation. Since what all participants want most of all is for the rugby game to take place, we can expect them to form two teams in the end.

We can also easily extend my reading group example. We can imagine that multiple members of the philosophy department want to form a reading group. Assume also that most but not all of them are needed for reaching the group size necessary to book a room. They all disagree on the reading material. But if joining a reading group with a sub-optimal book choice is preferable for enough of them compared with not establishing the reading group at all, then we can remain optimistic about the group's future.

According to Gardiner, three features characterise such extended battles of the sexes: First, partial cooperation suffices to make the rugby game or the reading group happen. If there are a few non-co-operators who do not join a team or settle on one of the

readings they liked less, it does not undermine the overall cooperation. Gardiner summarises this feature:

"(Partial Cooperation) There is a number M (such that M < N) which is the minimum number of players whose cooperation is necessary if some situation, which is dispreferred by all, is to be avoided." (Gardiner 2011: 89)

We only need so many cooperators for the rugby game or the reading group. If others spend time differently, that does not prevent the success of the rugby game or the reading group.

Second, in marginal cases where the participants are just short of enough people for the game/reading group, each prefers to cooperate over rugby or the reading group not taking place at all. In Gardiner's formulation:

"(Marginal Cooperation) If the number of others who are willing to cooperate is just short of M, then a given party prefers to cooperate, since each prefers to enjoy the benefits of cooperation and pay a share of the costs than to forego the benefits altogether." (Gardiner 2011: 89-90)

For example, if the reading group threatens to fail because only Bob and Jess are willing to compromise on the reading, other participants would become more likely[83] to compromise on the reading so that it can take place after all.

Third, none of the cooperators has any interest in disrupting the cooperation. No one wants to stop the rugby game or the reading group. As Gardiner summarises:

"(Passive Cooperation) Once the cooperating group is formed, the noncooperators prefer that it remain so, and succeed in its task. Hence, although they will not take on the costs

[83] The "more likely" does not mean to hint at a change of intrinsic preferences. These are presumed stable for modelling the situation as a battle of the sexes.

of cooperation, they will also refrain from disrupting the efforts of the cooperating group." (Gardiner 2011: 90)

Even if I refrain from exerting my energy in a rugby team, I stand back and let them play their game.

These three features sustain our optimism about multi-player battles of the sexes. We might hope that climate change negotiations also approximate these three features so that an agreement becomes likely. If only a limited coalition of states can solve the problem, if every state has an interest to join a coalition just short of the needed membership, and if the non-cooperating states do not interfere with the effort, we could become optimists about climate change negotiations.

While at first glance one might hope that the international situation has these features, Gardiner argues forcefully that it lacks them. Take partial cooperation: Not looking at the data, one might assume that the US, if it focussed on just this one goal, could stop climate change on its own without interfering in other countries. Then we could hope that a small coalition will take it upon themselves to avert catastrophic climate change. On this optimistic scenario, no country would stop all its emissions but a small number could make a considerable sacrifice which would suffice to save the climate.

But a small coalition does not suffice, since as Gardiner notes, "if either the Chinese or the Indians […] emitted like Americans, they would easily break the ceiling [of emissions before reaching potentially dangerous effects] all by themselves, even if the rest of the world cut its emissions down to nothing" (Gardiner 2011: 96). Gardiner (2011: 95-98) provides many more such data points, for example that if Bangladesh and Pakistan increased their per capita emission to the level of the US, then this would suffice to break

the ceiling.[84] The result is that the minimal cooperation coalition must include most states, at least insofar as they must not increase their emission levels too far over the current rate. The Vatican state can defect without problems, but nearly all major players must come together.

The next feature on the list is marginal cooperation, according to which participants prefer to join the coalition if it is just short of enough cooperators otherwise. We can combine the discussion of this feature with the discussion of the next: passive cooperation. Passive cooperation states that no one has an interest in undermining the cooperation. In the rugby and reading group examples these two features appeared reasonable enough. The participants want the game/reading group to happen and they have no incentive to undermine it.

Prima facie, one might also think that all states have an interest in the coalition being large enough to avert climate change since they might suffer the consequences of failure. But all states face incentives for defecting from a coalition of emission reduction and for effectively undermining the effort to avert climate change. While every state might profit from a successful coalition, all of them also derive an economic benefit from breaking the emission limits which grows with the size of the climate coalition. The more states join a climate coalition and stop using fossil fuels, the more the price of these energy sources decreases, so that others have a greater incentive to exploit them. If China and the EU cut their use of coal and oil, the prices fall and it becomes an even cheaper source of energy for the USA. The closer we get to a workable coalition, the greater the incentive to defect and improve one's own economy. Neither in the rugby nor the reading group example does the incentive to defect grow with the size of the cooperating coalition.

[84] The data on which Gardiner draws for the purpose of these calculations is currently being migrated to a new infrastructure, which can be found at: https://ess-dive.lbl.gov/ [27. 3. 2018].

The resulting increased emissions by defectors undermine the work of the coalition.[85] The USA could increase their oil and coal consumption because the prices decrease, and thereby counteract efforts by a small coalition trying to avert severe climate change. It is as if the more people joined the reading group, the more books from a limited number of philosophy books would go to those defecting from the reading group. The incentives increase for the defectors and, if they follow them, they undermine the cooperators.

The features which might make us hopeful regarding an extended battled of the sexes are missing in the case of climate negotiations. We should look for a less optimistic, but more plausible description.

Pessimistic Model: Prisoner's Dilemma

Pessimism, and in particular the assumptions that states face a prisoner's dilemma, dominates the current literature on climate change negotiations. As discussed in chapter two, in a prisoner's dilemma the agents prefer mutual cooperation to mutual defection, but also prefer personal defection to mutual cooperation. We can represent the game in the familiar matrix for two players:

PD	Cooperate	Defect
Cooperate	2, 2	0, 3
Defect	3, 0	1, 1

If we focus only on one player, we can represent the dilemma as follows:

[85] For a longer discussion see Gardiner 2011: 98-101.

PD	Other Cooperates	Other Defects
Cooperate	R(eward)	S(ucker)
Defect	T(emptation)	P(unishment)

Assuming T>R>P>S,[86] there is a reward for mutual cooperation, but everyone fears being the sucker and faces temptation. If both fail to cooperate, they receive punishment relative to the outcome of mutual cooperation.

Given this structure, we might say with Gardiner that it is collectively rational to cooperate and individually rational to defect. The participants prefer the outcome of mutual cooperation over mutual defection, but regarding their own decision they always prefer to defect (cf. Gardiner 2011: 104). This divergence of collective and individual rationality makes the situation a dilemma. The outcome is sub-optimal for all agents, even though individually they act rationally.

Many have been tempted to model climate change negotiations as such a prisoner's dilemma.[87] A state, say the USA, prefers mutual cooperation to mutual defection in curbing greenhouse gas emission. However, it also prefers for the others to cooperate and itself to defect so that climate change would be averted without the USA having to pay the cost. In such an interpretation, the USA find themselves in a prisoner's dilemma with other states. Although to cooperate would be collectively rational for the participants, individually, defection remains rational for the USA.

[86] I take this representation from the influential Axelrod & Hamilton 1981.
[87] See Gardiner 2004: 594-595. See also Brennan 2009.

The structure fits with the pessimistic conclusion we drew from the discussion of the battle of the sexes. We need cooperation from almost all participants, and defectors undermine the aims of cooperation. The agents have no incentive to join the cooperators if the climate coalition falls just short of the needed number, instead everyone has an incentive to defect. Gardiner's data and arguments speak in favour of the prisoner's dilemma over the battle of the sexes as a basis for modelling climate change negotiations, but important complications are missing from this well-known game.

Complications

Some authors have a stubborn attachment to thinking of climate change as a prisoner's dilemma (for example Brennan 2009), but a little reflection reveals that such a model remains too simplistic. Any rational choice theorist worth her salt can spot questionable assumptions needed for modelling climate change negotiations as a prisoner's dilemma.

Consider the following complications:

Multiple Agents

More than two players are involved in the negotiations. There are over a hundred of them if we simplify and only count the states involved in climate negotiations. Counting all relevant agents influencing the negotiations, for example NGOs and corporations, the number increases further. The states must form a large coalition to deflect the negative consequences of climate change. Cooperation becomes more difficult once this complication is introduced.

Multiple Rounds

Climate negotiations are not a one-shot game, but rather take place repeatedly (cf. Madani 2013: 70-71). Before the Paris Summit, delegations met in Copenhagen and diplomats also stayed in contact between these meetings. In principle, repeated prisoner's dilemmas are more likely to result in cooperation (as I mentioned in chapter three, see also Axelrod & Hamilton 1981). Other players can punish defectors by themselves defecting in consequent rounds.

Multiple Decisions

Not only are the negotiations repeated, but during each round "[e]ach agent must make multiple decisions" (Gardiner 2011: 107). Each state decides on how it regulates pollution on a variety of dimensions. The multiplicity of decisions renders every meaningful agreement more complex and therefore more difficult to achieve.

Sanctions and Side Payments

Players can engage in sanctions and side payments. Many expect the richer countries to support the less wealthy countries financially or with technological transfer. States who do not support efforts to combat climate change can be punished in the context of other international affairs, for example in trade negotiations or even war. These options help mitigate the incentives to defect (cf. DeCanio & Fremstad 2013: 180-181).

Degrees of Cooperation

Cooperation and defection come in degrees rather than in disjunctive categories. States can reduce their emission of pollution in degrees and climate change comes in degrees (for a short discussion, see Wood 2011: 156). We do not find only cooperators and defectors, but a spectrum of partial cooperators.

Degrees of Significance

Cooperation and defection matter to varying degrees depending on the state doing so. That the USA defected from the Kyoto protocol reduced its effectiveness more than if Austria had defected. The model should take the relative importance of the players into account.

Solution Concepts

Typically, papers on climate change negotiations assume the Nash equilibrium as a solution concept, where a solution concept is the rule for determining the how the game will be played. In the case of a strict prisoner's dilemma, this doesn't matter since the Nash equilibrium state of mutual defection is also dominant. However, if we introduce the previous complexities, it is possible that no dominant option remains, which makes the choice of a solution concept an important factor. DeCanio and Fremstad (2013: 179) suggest that players might be risk-averse and therefore follow more of a maximin solution concept in which they maximise the minimal pay-off. Madani (2013) criticises DeCanio and Fremstad for not taking even more solution concepts into account.

This list of complications remains incomplete[88] and the decision as to which of the complicating factors we must include to achieve a predictive model is a controversial one. The changes have highly complex and interrelated consequences for the prospect of a climate change agreement. While the repeated and gradual nature of climate negotiations might increase our hope for an agreement, conversely, the large number of agents needed for success reduces it. One way or another, we have to change the model and leave the charming simplicity of the prisoner's dilemma behind to move toward a plausible representation of actual climate negotiations.

Gardiner takes some complications into account and argues that the climate situation resembles Hardin's classic discussion of the tragedy of the commons more than a standard prisoner's dilemma (cf. Gardiner 2011: 108-114). In Hardin's example, the herdsmen of a village share a common piece of land on which their livestock can graze. All of them have an incentive to add further livestock to their herd until the common grazing grounds become exhausted (cf. Hardin 2009). Herdsmen who hold back and do not enlarge their herd also end up with an exhausted common piece of land without having reaped the benefits of more livestock.

While the tragedy of the commons is often understood in terms of a prisoner's dilemma, it diverges in multiple points. The herdsmen might not add all the livestock at once but instead do so gradually and incrementally. Likewise, countries do not emit pollution all at once but instead emit a given quantity over time. In contrast to a prisoner's dilemma, no single decision to defect is decisive. In each round, a decision to defect "*erodes* the collective good—making the full collectively rational outcome unattainable—but does not make further cooperative efforts pointless" (Gardiner 2011: 110). Even if the USA

[88] One often mentioned complication I leave aside is the intergenerational nature of climate change (see Gardiner 2003, 2011). Partially this gap results from looking at group agents: although their members change, the group agents remains the same. A further theory of how generational change affects group agents would be needed.

defects during some rounds of negotiations, such as the Kyoto protocol, the game does not end. Other countries can still reduce emissions and if the USA re-joined later the new cooperation would still mitigate the consequences of climate change.

The classic tragedy of the commons example also includes the other aforementioned complications: We could argue that the decision in Hardin's example and in climate change negotiations comes in degrees. Just as the herdsmen can add none or varying numbers of livestock, states can cut all or varying fractions of their emissions. In sum, the prisoner's dilemma serves as a superior model of climate change than the battle of the sexes, and the tragedy of the commons adds important complications on top of the prisoner's dilemma.

That the world might face a situation resembling a tragedy of the commons is a worrying prospect indeed. Hardin called it a tragedy for a reason. As in the case of the prisoner's dilemma, collective and individual rationality diverge (cf. Gardiner 2011: 108-109): Each herdsman prefers that the common is sustained by mutual cooperation over it being destroyed by mutual overgrazing, but in each round, everyone individually prefers to defect rather than to cooperate. The same might apply to climate change negotiations. Adding multiple rounds and degrees of cooperation and defection alone does not end the tragic structure, although it might alleviate the problem somewhat.

In recent decades, Elinor Ostrom and others following her pioneering research have shown that human groups can overcome such tragedies (cf. Ostrom 2015). From Swiss and Japanese peasants sharing meadows and forests, to Turkish inshore fisheries, to Spanish and Philippine farmers managing complex irrigation systems, human groups have successfully shared commons for long periods without tragic results.

However, these positive cases rely on clearly defined boundaries of the commons, strong social cohesion, and the possibility of excluding exploitative defectors. Gardiner provides

reasons to believe that the conditions for overcoming the tragedy do not exist in the case of climate change:

"From an international perspective, social capital is weak, not everyone supports regulation, excluding noncooperators from emitting carbon is very difficult (if not impossible), emissions are difficult to monitor, and the rate of change in emissions in at least some economies is considerable (e.g., China is building new power plants every month)." (Gardiner 2011: 116)

The case of climate change lacks the features Ostrom and her followers rely upon that would provide more hope about the tragedy of the commons. The outlook for climate change negotiations offered by the rational choice models remains dismal. Even taking significant complications into account, we stay closer to the pessimism of the prisoner's dilemma than the optimism of a battle of the sexes. But the resources of standard rational choice theory are not yet exhausted.

Turning the Models on Their Heads: Implementation Theory

Following Gardiner, we assumed that states are confronted by some type of game in climate change negotiations, and we then attempted to find out how best to model such negotiations. But we can turn rational choice theory on its head and ask: How should we organise climate change negotiations to avoid a situation in which collective and individual rationality as described by Gardiner diverge?

Implementation theory is a field that asks how non-cooperative games can be designed "so that their solution […] corresponds to a social optimal outcome" (Wood 2011: 164). Implementation theory typically assumes that the preferences of the agents determine the social optimum, which is roughly what Gardiner considers the collectively rational

outcome. Such social optimality can fall short of a philosopher's loftier requirements of justice, but it improves the situation compared with a prisoner's dilemma ending in mutual defection where each participant could increase its pay-off if they only mutually cooperated. If states face a prisoner's dilemma in climate negotiations and defect, then all of them could suffer less from climate change without anyone being worse off.

Boadway et al. (2011) offer one exemplary mechanism designed to achieve socially optimal climate change negotiations (also discussed in Wood 2011: 166-167). In this case, the optimality concept is Pareto-optimality. In an example with two countries, this means that we hold the benefits of one country at least constant while maximising those of the other (cf. Boadway et al. 354).

I introduce Boadway et al.'s model informally and present the consequences without going into the mathematical details. The main purpose is to explore to what extent standard rational choice models that assume stable intrinsic preferences can help to improve the outcomes of climate change negotiations.

The Boadway et al. model assumes that states can conditionally commit themselves to abatement levels of pollution. We do not necessarily need a central enforcing authority, but in one way or another the USA and all other states must be able to credibly commit themselves to a reduction of their greenhouse gas emissions.

Given this assumption, Boadway et al.'s model has the following consecutive steps:

1. Each state simultaneously commits to matching rates for all other countries' direct abatement levels.
2. Each state simultaneously commits to its direct abatement levels.

First the participants, that is the USA, China, the European Union, and so on, commit to certain matching rates. For example, the EU might commit to cut one and a half tons of

CO_2 for every ton India cuts. In the second step, each member declares its own direct abatement levels which the others must now match. India specifies the number of tons it will cut and the EU has to match them in addition to its direct abatement. With these two steps, each state commits the sum of its direct abatement level plus the direct abatement levels of other states multiplied with the matching rate for this other state.

The outcome of this negotiation procedure is socially optimal and the effective abatement costs faced by the states are analogous to the Lindahl price, that is, the price of abatement varies according to the satisfaction of the preferences derived (cf. Boadway et al. 2010: 357-358). The optimality implies that individual and collective rationality no longer diverge as Gardiner sketched, since no one can be better off without another player being worse off.

As an extension, Boadway et al. add a third step to the model:

3. States engage in trading of their emissions quotas.

This additional step helps to equalise the marginal benefits of emissions across all states. The benefits for an extra unit of emission become the same for all participants. Boadway et al. also provide further refinements (Boadway et al. 2010: 360-364), but for our purpose this mechanism will do.

The outcome of this negotiation model is superior to what we expected after Gardiner's discussions: No longer do the agents end with mutual defection although they could profit more from mutual cooperation. The mechanism illustrates how standard rational choice theory can provide action guidance for climate change negotiations. The Boadway et al. model suggests a way to set up negotiations that improves upon the mutual defection of a multi-player prisoner's dilemma. The action guidance, however, also has its limits.

One limit lies in the notion of social optimality. As already suggested above, a Pareto-optimal outcome might fall short of the requirements of moral and political philosophers. Even a strict utilitarian remains unsatisfied, because the Boadway et al. model only considers the preferences of the participating players, that is, the states. Gardiner (2011) argues that one shortcoming of current attempts to address climate change is a neglect of future generations. Current group agents too strongly discount future generations, insofar as neither the USA nor any other states appropriately represent the interests of its population in a hundred years. But if the states discount their own future generation too much then the outcome of the Boadway et al. mechanism fails to efficiently take the preferences of future generations into account.

But this point still grants that the model's assumptions hold and that we can implement it, which we have good reasons to doubt. The model assumes that states have the capacity to credibly commit themselves to matching rates and direct abatement levels. One might suggest that sovereign states are always able to back out on their commitments and find it hard to commit themselves credibly.

The case of the USA illustrates the difficulty: For a credible long-term commitment, the President needs the support of Congress, which proved difficult to get during the Kyoto negotiations. The Paris summit negotiators tried to avoid any formal commitments, which would have required an official decision by Congress, because it had proven exceedingly difficult to gain such support.[89] That, of course made, it easy for a President Trump to walk away from commitments Obama had tried to put in place for the USA. The group agents in this scenario find it difficult to commit in the ways needed for Boadway et al.'s model to work.

[89] See: https://www.washingtonpost.com/news/powerpost/wp/2015/11/30/trick-or-treaty-the-legal-question-hanging-over-the-paris-climate-change-conference/ [29. 7. 2017]

The instability of commitment might reduce the efficiency of the outcome. For example, the Obama administration might have tried to have the USA commit to smaller emissions than would have been strictly efficient, because they hoped that such a commitment had a better chance of withstanding following administrations. The limits on commitment undermine the positive results of the Boadway et al. model.

In addition, all implementation theory models raise the worry of second-order games. How can we establish a general mode of negotiations? It is not Boadway and his team who decide how climate change negotiations proceed but the states themselves. Since the Boadway et al. mechanism promises to overcome a divergence between individual and collective rationality, we can hope that the states are enlightened enough to follow their own interests and endorse the two-step negotiations. Assuming the mechanism works, all agents should be better off following the procedure and no one worse off.

However, establishing such a negotiation mechanism has side effects which might keep states from accepting it. States might worry that putting such an international framework in place might threaten their interests in other respects. They might believe that there is more to gain in the international arena through conflict and keeping negotiations minimal. Then states face incentives to avoid Boadway et al.'s mechanism of negotiations. In other words, the action guidance of implementation theory depends on our capacity to impose a structure on climate negotiations, which might prove limited.

Furthermore, and most importantly for my purposes, standard rational choice models assume fixed intrinsic preferences. But agents change their preferences after the negotiations or during them – this can undermine the efficiency of the outcome. Consider the following scenario: India commits to certain matching rates and a direct emission abatement level in two steps together with all other states. However, afterwards India becomes more environmentally aware. Now the efficient outcome would include a

higher abatement level. Of course, India might reduce its abatement even more than it is committed to, but since the negotiations are over, the other agents might not match this additional abatement.

To summarise, Boadway et al.'s model provides guidance on how to conduct climate change negotiations in a way that increases the chances of an efficient agreement. But the approach has shortcomings.

First, it is dubious whether a Pareto-efficient outcome according to prior intrinsic preferences meets our normative requirements. It might diverge from a just result. What if the preferences of the states make a major climate change the efficient outcome because they discount future generations too much?

Second, the assumptions of the models remain debatable. The extent to which states can commit themselves to levels of abatement is called into question, the threat of second-order games looms, and the change of preferences undermines the efficiency of the result.

Implementation theory is a valuable tool for avoiding catastrophic climate change. As such, we should include it in our toolbox, but we should not limit ourselves to such models. Particularly if states overly discount future generations, we should look for solutions that include preference change. So, we turn to the Negotiated Order approach.

The Negotiated Order Approach to Climate Negotiations

The Negotiated Order approach as I have reconstructed introduces an account of intrinsic preference change and will thereby allow us to see more options for climate change negotiations.[90] However, the approach also has a shortcoming: In contrast to standard

[90] The work within the literature on international relations that bears the greatest resemblance to my proposal is probably Alexander Wendt's (1994) "Collective Identity Formation and the

rational choice theory, it does not provide sophisticated mathematical models. Blumer and Strauss certainly did not quantify their theories and, so far, I have hardly improved upon that. As a result, I cannot offer as specific a picture as standard rational choice theory. I cannot provide quantitative predictions of the consequences of certain mechanisms, nor proofs that outcomes would be Pareto-optimal.

However, we can hope to develop mathematical models based on the Negotiated Order perspective. After all, Cohen and Axelrod managed to formalise their model of adaptive utilities. The appendix on Dewey's decision theory takes the first steps towards a formalisation of Dewey's claims. While the limitation of the Negotiated Order is contingent and temporary, for now it remains a deficiency.

The shortcoming results from giving up the assumption of stable intrinsic preferences. This assumption helped us to construct a mathematical model, but it also obfuscates available options. The Negotiated Order approach lacks mathematical models so far, but it does uncover these options. I pointed out that the Boadway et al. model runs into problems when agents change their preferences. In addition, we hope for preference change on the assumption, argued for by Gardiner (2011), that current climate negotiators inordinately discount future generations. To discuss how we might face these issues, we need to include preference change in our theory.

For the reopening of possibilities to be a productive venture we need some guidelines for preference change. A model that includes arbitrary, unlimited preference change has little predictive power and no plausibility. Anything could happen. In chapters two and three I discussed the limits the Negotiated Order approach puts on intrinsic preference

International state". He too draws on the tradition of symbolic interactionism and its affinity to preference change. However, he focuses on a different mechanism of preference change bound up with identity instead of using the Negotiated Order approach. As a result, our proposals diverge considerably.

change in a general and largely informal way. Agents do not change their intrinsic preferences arbitrarily, but in response to Problematic Situations.

The features of the Problematic Situation and the events during the exploratory phase influence the direction of preference change. As discussed in chapter three, the agents engage in signalling during the exploratory phase and this affects what motivational profile finally settles in. The agents can align during parallel processes of preference change. To see the effects this preference has on modelling climate negotiations, consider the following informal example, which provides only the first taste of what is to come, and makes assumptions that I will later question.

For a start, assume a tragedy of the commons depiction of climate change negotiations. We can think of it as a more complicated version of a repeated prisoner's dilemma in which an overwhelming number of participants must cooperate for a successful climate agreement and the more who cooperate, the higher the incentives for the defectors.

In an infinitely repeated prisoner's dilemma between two players, we would have reason to hope that they settle on cooperation after a while. But various features of the situation undermine our optimism: a large proportion of agents needs to cooperate. The more agents cooperate, the more the prices for fossil fuels decrease for the defectors, incentivising their use. Internal controls remain weak and cooperators cannot exclude defectors from the commons, that is, the world climate.

We can expect that, round for round, important agents defect because they do not want to be the suckers in this round. For example, the USA prefers for others to cooperate during each round of climate change negotiations, but for themselves to defect, and mutual defection over being a sucker, while preferring mutual cooperation by a larger coalition over general defection.

We can illustrate the situation *in a single round* with a typical 2x2 matrix:

	Others Cooperate	Others Defect
Cooperate	R	S
Defect	T	P

Assuming again T(emptation) > R(eward) > P(unishment) > S(ucker).

To clarify that we do not actually have two agents here as we would have in a standard prisoner's dilemma, I labelled the columns "*Others* Cooperate/Defect" rather than just "Cooperate" and "Defect". The USA can choose between two actions, the pay-offs of which depend on the choices made by multiple others.

It is common knowledge that a large coalition is necessary to avoid severe climate change. This knowledge informs each player's choice between the possible acts, reducing the likelihood of cooperation. We do not have to change anything about the outlined pay-off matrix, because the pay-offs remain the same. But we should keep in mind that each player has information reducing the probability of the "Others Cooperate" option and therefore the *expected pay-off* of cooperating oneself.

Up to now, we stuck to standard rational choice theory, but the failure in finding an agreement that limits the increase in global temperature is likely to lead to Problematic Situations. The consequences of dramatic climate change cause disruptions and agents will identify problems. Droughts, floods, rising sea levels, and similar events disrupt economic activity and plans. Having identified problems, the agents show an increased likelihood of intrinsic preference change. The Negotiated Order approach predicts that they will open up to preference change.

The USA might open up to becoming a more cooperative player.[91] It tries out a new tentative preference for preventing climate change. To give a toy example, this preference might result in the following transformation of the prisoner's dilemma for the USA:

	Others Cooperate	Others Defect
Cooperate	R+i	S+i
Defect	T	P

In this illustrative example, an i-term is introduced which changes the dynamics of the game. If $i \geq (T-R)$ then the game is no longer a prisoner's dilemma because now temptation is eliminated. If $i > (T-R)$ and $i > (P-S)$, cooperation becomes the dominant strategy for the USA. To put it informally, cooperation becomes dominant if the new preference for cooperation outweighs the prior pay-off difference between cooperating and defecting.

One might fear that the USA would only end up in a worse position if it underwent such a preference change. The USA has to pay the price of cooperation, for example, a reduction in economic growth resulting from emission regulation. Their cooperation does not suffice to achieve a significant climate agreement if all the other participants stick with their strategies. The agent with the preference for cooperation could turn into a continual sucker. Because the USA would experience the new cooperation as positive, it would not see itself that way, but from the outside it would look as if the USA were simply

[91] An intrinsic preference for a climate agreement is not the only option. If the states developed a stronger intrinsic preference for preserving nature, this would have a similar effect.

making sacrifices without receiving any reward. Such preference change might appear unlikely and unwelcome.

The Negotiated Order approach, however, tells a more complex story. At first the USA only explore this new preference for cooperation without yet settling on it. They hold it in a qualified tentative mode. They can signal this potential change, but if they do not experience others joining in, they might not change their preferences permanently. The exploratory phase proves once more an important contribution of the Negotiated Order approach.

The signalling of a potential preference change in favour of cooperation can take various forms. For example, after opening up to preference change, the USA might cooperate in the next round of climate negotiation rather than defect. Even knowing that in this round no one will join in the cooperation effort, for this round the agent foregoes the benefits accruing to defectors. This action signals to other agents a tendency of the USA to become more cooperative. Given the common knowledge of the situation and the previous behaviour of the USA, the cooperation act increases the probability of the USA to develop a more cooperative motivational profile. Furthermore, the signal is relatively trustworthy, because the USA pays for it with concessions in this round of negotiations. It is a case of costly signalling.

If the USA defected again and again during repeated climate change negotiations, and then started to unilaterally cooperate, other participants would interpret the move as "taking leadership" and "standing up for the value of international cooperation". This signal might affect the preference changes of other states, which also encountered Problematic Situations and entered exploratory phases. The participants might undergo preference change in a coordinated manner, thereby avoiding turning into perpetual suckers.

If enough agents undergo such a coordinated change of motivation, they can form a coalition despite the dismal starting situation. Intrinsic preference change and signalling during the exploratory phase allow us to break out of the tragedy of the commons. They do not guarantee it, but they do increase the chance and provide at least a glimmer of hope.

This kind of more optimistic scenario becomes visible from the Negotiated Order perspective, while it is not a potential outcome for standard rational choice models. With a different theory of motivation, a different way to cooperate opens up. But the real story is more complicated than the example.

Complications

The example I gave described a simplified scenario. As for standard rational choice models, we might consider various complications to make the description more accurate. I list the complications for a standard one-shot prisoner's dilemma again because they have different consequences assuming the Negotiated Order approach. Some of these complications were already included in the example, some are added to it:

Multiple Agents

I included the assumption that the negotiations involve multiple players in my example, but one can make various assumptions about how many agents negotiated and how many are needed to form a coalition that will lead to success. The higher the number of agents needed for success, the unlikelier the success. This holds for standard rational choice models and for my example. The more agents are involved, the smaller the likelihood that enough of them converge in their preference change. We might end up with a coalition too small to fend off catastrophic climate change.

Multiple Rounds

Rather than being a one-shot game, climate change negotiations are a recurring event. I assume repeated games already in my illustrative example. The repetition provides the room for exploratory phases and signalling among the states. The repetition also allows punishment of those who give wrong signals, and therefore supports truthfulness. This complication increases the chance of cooperation for standard rational choice models and it increases the chances for aligning in preference change for Negotiated Order models.

Multiple Decisions

During each round, agents make multiple decisions rather than just one. Each state decides how to regulate pollution in a variety of dimensions. The consequences of this complication are difficult to predict. On the one hand, the multiplicity of decisions makes it more difficult to align on all relevant preferences and find an agreement. On the other hand, the variety of decisions allows for fine-tuned signalling, potentially increasing the chance of cooperation. Without a detailed model, few conclusions can be drawn.

Sanctions and Side Payments

If players can engage in sanctions and side payments, the chance of cooperation under the assumption of stable preferences increases. Those who defect can be punished and those who cooperate rewarded. With preference change entering the picture, sanctions and side payments potentially play the further role of influencing the direction of preference change.

Agents who try out new preferences can be rewarded or punished for their direction of preference change. Given that agents respond to such contextual cues, as the Negotiated Order approach assumes they will, the feature increases the chance of cooperation even further than in the standard model.

Degrees of Cooperation

Cooperation and defection come in degrees rather than as disjunctive categories, as do emissions and climate change. By virtue of this feature of climate negotiations, preference change does not have to be dramatic to make a difference. A slight preference change in favour of cooperation might already make a difference to the outcome. The USA might become willing to accept slightly lower emissions. If enough states undergo such slight preference change, it might have an important overall effect. Even if our hope for avoiding climate change altogether should fail, the Negotiated Order approach might still reveal a way to avoid the most destructive scenarios.

Degrees of Significance

Although only a large coalition can avoid climate change altogether, agents still matter to different degrees. For example, the cooperation and defection of the USA has more impact than Austria's cooperation and defection. Because cooperation and defection of different states matter to varying degrees, the preference change of these states matters to varying degrees. A preference change on the part of the USA has greater significance than would a preference revision undergone by Austria. Accordingly, even if not all states fulfilled the Negotiated Order account of motivational change, it could still matter as long as it applied to relevant agents in the negotiations.

Multiple Solution Concepts

Multiple solution concepts other than the Nash equilibrium are also worth considering. Depending on which solution concept agents follow, different changes of intrinsic preference can lead to a success in agreements. Again, to become more specific we would need a detailed quantitative model.

As we can see, the complexities we can include in standard rational choice models also affect the picture offered by the Negotiated Order approach. In addition, new complications arise that did not come into view for standard rational choice theory:

The Agents' Propensities to Change

Participants, in this case states, differ in their inner structure, making it more or less likely that they undergo preference change in response to Problematic Situations and align with other agents. For example, Germany might be more likely to change its intrinsic preferences than a country in which the ruling elite is more heavily invested in the extraction of fossil fuels. Assuming agents are bogged down in a situation resembling a prisoner's dilemma or a tragedy of the commons, we want them to be more likely to undergo preference change so as to overcome the dilemma.

Contextual Support

A context can offer more or less room for exploring potential new preference and signalling. In this regard, it might affect whether agents align during preference change.

In the case of climate negotiations, the international system provides the context. It determines a number of significant variables, for example how easy it is to have another round of negotiations, or the costs of signalling.

Speaking generally and without a detailed model to prop the claim up, if the context leaves ample room for signalling, this might improve the chances of cooperation. Consider the following example: The USA might be willing to try out a new preference but it would need support from other states to settle on it. If the participants do not have enough room for signalling and the USA are unsure whether others might change their preferences too, cooperation becomes less likely.[92]

These complexities call for further investigations, consisting partially in mathematical model building, partially in empirical research. In effect, I offer a whole research programme for the Negotiated Order approach. Using my suggestions about how to reconstruct the Negotiated Order approach and applying it to group agents, new inquiries into climate negotiations become available. This kind of research programme goes beyond the boundaries of this thesis and requires further social scientific expertise. Instead I want to explore how we can use the Negotiated Order perspective for action guidance, just as implementation theory employs standard rational choice model for action guidance.

Assuming certain ends, for example a climate deal limiting emissions further than what has been achieved so far, we can draw suggestions from my reconstructed Negotiated Order approach. To give a general example, if we could make an increased preference for

[92] There is also the worry that states might manipulate how problematic a situation appears for other each other. My model raises awareness of such options. I thank Yonatan Shemmer for raising this issue.

cooperation salient for a state in response to a Problematic Situation, this would increase the likelihood of a successful agreement. We can draw various similar proposals from the Negotiated Order approach. However, before I turn to the details of such a proposal, I face a fundamental objection: Is the relevant preference change even plausible?

Objection: The Plausibility of Relevant Preference Change

Even accepting that agents, including group agents, can undergo preference change in response to Problematic Situations, one might consider such change limited. A country might undergo change to some of its motivational states, but will the change ever reduce its interest in power and economic wealth over international cooperation? Some preferences might be practically unchangeable. The potential suggestions offered by the Negotiated Order perspective could turn out to be uninteresting if these limits are too strict.

In the context of climate change, we find various positions that endorse strict limitations on the preferences of states. The school of Realism in International Relations (cf. Grundig et al. 2001) claims that states always look for relative power gains. Assuming this were true, states might still undergo some preference change, but the resulting preferences would only become relevant if the choice had no effect on the relative power. For example, if a state could win the same relative power either by going to war or by entering a peace treaty, then a preference for peace might decide. Such scenarios are rare and we should not expect them in the case of climate change. The results of climate negotiations will likely have consequences for the relative power of states. Accordingly, this sort of Realism imposes limits on preference change that would render the Negotiated Order perspective uninteresting.

For another example, consider Geoffrey Brennan's (2009) charmingly dismal paper "Climate Change: A Rational Choice Politics View", which is predicated upon the assumption that one can model climate change as a multi-player prisoner's dilemma. While global cooperation might increase overall utility, we should have little hope for achieving it according to Brennan. The preferences of states do not allow it and any hope that this might change seems naïve. He claims that defecting from any climate cooperation is in the national interest of any nation, for example Australia. In response to anyone who disputes this position, he writes "in my view, the claim itself [that it is not in Australia's national interest to reduce carbon emissions] is uncontestable. And people who dispute it simply do not understand the nature of the problem!" (Brennan 2009: 313)

Brennan probably wants to claim that Australia's unwillingness to sacrifice its economic well-being for a preference to reduce climate change is beyond reasonable doubt. That Australia would change its preferences in a way that puts cooperation over its economic well-being is just implausible, or so Brennan believes.

In Brennan's picture, we face a limit to preference change that appears to threaten the usefulness of suggestions from the Negotiated Order perspective. If preference change always fails to establish a preference of cooperation over economic well-being, the states remain in a game resembling a prisoner's dilemma. At least that is what Brennan suggests.

But simply asserting that some preferences remain unchanged and are therefore exogenous to the negotiations is not so much an objection against my account as begging the question. While a state's preference for power is more deeply engrained than other preferences, more needs to be said to establish that the preferences relevant for climate change negotiations are practically stable. In the end, nothing but difficult and empirically informed research settles the question as to which preferences lie beyond change.

I cannot undertake such a far-reaching investigation here. However, one principled consideration speaks against endorsing too dismal a picture of preference change in the case of climate negotiations. Relative power, economic wealth, climate change, and emission reduction all come in degrees. Consider a case in which the USA could pass a regulation that avoids flooding caused by rising sea level and leads to an overall decrease of real GDP growth by just one single dollar. Is it plausible that the USA would not accept losing this quantity of relative power (or economic wealth) to avoid the flooding? I venture that such a proposal is implausible.

Relative power and economic wealth do not always trump avoiding the consequences of climate change, because the consequences of climate change can be huge and the costs in terms of power and wealth negligible. Granting this, one can change my example incrementally. Would the USA accept losing a real GDP growth of a $100 to avoid damage by flooding? What if the damage occurred in a partner country rather than in the USA itself? I hope the USA would bear this cost. Admittedly, at some point the preferences for the climate over power and wealth break down. We would have to determine the exact point empirically. But I see no reason why it should be unchangeable.

Usually authors simply make the variables categorical rather than a matter of degree: Brennan considers cooperation versus defection. With the assumption of categorical variables, it becomes more plausible that power and wealth always trump cooperation and concern for the environment. But the assumption of categoricity simplifies the situation.

I made a similar simplification above when I discussed climate change negotiations as resembling a repeated prisoner's dilemma. I offered a toy model in which the dilemma is transformed to the following game:

	Others Cooperate	Others Defect
Cooperate	R+i	S+i
Defect	T	P

We should not confuse the simplified model with the actual negotiations. Our contribution to climate change, our willingness to cooperate in battling it, and climate change itself, all come in degrees. Shifting the degrees can already have a significant impact, depending on the number of states undergoing such a change.

Even if preference change does not lead states to become whole-hearted cooperators sacrificing power and wealth for the common good, changes at the margin can help. The complication of degrees of cooperation unmasks Brennan's perspective as too dismal. Preference change can have an effect at the margin of cooperation. Therefore, we should look at proposals of how we can contribute to positive preference change.

Proposals

I assume that we aim at increasing the probability of an agreement on a significant emission reduction. Given this aim, the Negotiated Order approach provides proposals for action, which I turn to now. For the sake of illustration, I assume that the agents face a multi-player prisoner's dilemma as sketched in my original example. We can again use the following matrix for illustration purposes:

	Others Cooperate	Others Defect
Cooperate	R	S
Defect	T	P

As mentioned, this modelling simplifies the issue since decisions in climate change negotiations come in degrees. But most of the proposal remains transferable to more complex models. A move that increases the change for more cooperative preferences applies whether cooperation comes in degrees or as a disjunctive category.

My proposals remain informal. I also leave open whom the proposal addresses. Boadway et al. do not explain who should establish their negotiation mechanism, and I do not explain who should implement the changes I propose. Instead of addressing any particular group, I use a general 'we', that is, we should do such-and-such to avoid severe climate change. My discussion focusses solely on the wheels we − whomever that may refer to − can turn to achieve cooperation in reducing climate change. I aim to show that the Negotiated Order approach to negotiating group agents makes a significant difference. For that purpose, I allow myself to stay at a rather general level and to make implicit assumptions about causal relations.

We can distinguish two types of proposals resulting from the Negotiated Order approach to climate change negotiations: proposals as to how to change group agents and proposals as to how to change their context. In effect, the two new complications mentioned above serve as the starting point for my discussion. I go through both types of proposals and discuss how we could increase the probability of reducing emissions.

Proposals for Reshaping Group Agents

Here is an optimistic scenario: Unexpected events disrupt the activity of agents, they identify a problem, and open up for preference change. They enter an exploratory phase during which they can signal with other agents to align in their preference change. In response, the agents align on new, more cooperative preferences so that they can form an effective coalition. We can reshape the agents in regard to every of these points of the process.

In our illustrative toy example, the motivational changes of the group agents are supposed to transform the original matrix given above to the matrix including the additional variable i:[93]

	Others Cooperate	Others Defect
Cooperate	R+i	S+i
Defect	T	P

So that i > (T-R) and i > (P-S), as many agents as possible should undergo such a change.

We want every agent to be such that it is more likely to

1. undergo a transformation of its pay-offs, in particular,
2. introduce an i such that i > (T-R) and i > (P-S) during the exploratory phase, and
3. signal the potential preference change and settle on it if other agents signal a similar tendency.

[93] Again, the idea of a simple preference change, which leads to the introduction of the i-term, only serves the purpose of illustrating the consequences of endorsing the Negotiated Order approach. A more appropriate model of actual climate change negotiations would be more complicated and the proposals would have to be adapted.

For example, we want the USA to be more likely to undergo a transformation of preferences: in particular, a pro-cooperative change that outweighs the prior incentive to defect. Furthermore, the USA should signal this potential preference change to the other states and settle on it if they truthfully signal back a tendency for a similar change. If enough agents undergo such change, then they overcome the situation resembling a prisoner's dilemma. We want to increase the probability of a coordinated preference change in favour of far reaching climate cooperation. How can we achieve this?

According to the Negotiated Order approach, an agent is more likely to undergo a motivational change that transforms the game if it identifies a problem in response to a disruption of activity. Whatever we can do to increase the likelihood of this kind of problem identification contributes to the success of the scenario. For example, we might assume that a state with a sustained public discourse is more likely to identify a problem in response to a disruption and an indeterminate situation. Then we should support such a public discourse. Incentives for ignoring disruptions decrease the chance of a game transformation. We should work on removing such incentives. Again, the exact actions we should take depend on the circumstance. Without empirical research, we can only guess what the causal connections could be. However, the Negotiated Order approach tells us what to look for and what to do when we have found it.

Not only should the agent identify a problem, but its structure should also be flexible enough to allow for preference change in the first place. Otherwise the group sticks to its preferences even after identifying a problem. Political systems can make it easy or difficult to introduce a change of policies. For example, if the foreign policy is completely governed by a multi-chamber parliamentarian process, then a policy change will prove to be difficult.

This point suggests that we want change to be easy to achieve. On the other hand, only if the state achieves some consistency in its policies will it function as an agent at all.[94] The states need to be agents capable of credibly committing themselves to certain emission abatement levels. If a commitment can only be revoked by a multi-chamber parliamentarian process, then it would be more credible. We should be aware of trade-offs. The considerations of the Negotiated Order approach put weight on one side of the scale, the side of flexibility, but there is also another side to it. Again, the final advice depends on empirical circumstances which we should include in our models.

Being more likely to change one's preference is not everything. The preference change also should point in the right direction. If the i variable is negative, then the chances of a cooperative result decrease rather than increase with preference change. States should be such that they are likely to try being more cooperative in response to Problematic Situations. Which changes within a state lead to a propensity for becoming cooperative depends, once more, on the concrete case. In the USA, but elsewhere as well, strengthening the influence the scientific community has on policies might be advisable. To give perhaps the least innovative piece of advice resulting from my approach, voting for appropriate parties increases the propensity of the right preference change.

Not just one state should undergo a pro-cooperative preference change. We hope for a coalition of states that align on more cooperative intrinsic preferences during an exploratory phase. Signalling helps us to achieve this goal, as I discussed in chapter three. The states need the capacity to engage in signalling in order to align their preference change and reach a strong agreement.

For example, enhancing the state's diplomatic capacities so that it can send trustworthy signals eases the alignment of preference change processes. Strengthening these

[94] According to the functionalism discussed in chapter five.

capacities forms an important contribution to approaching the positive scenario I sketched. Of course, diplomacy has been praised before. The Negotiated Order approach contributes the insight that signalling also affects the preference change and therefore matters for achieving further-reaching cooperation. It offers a further consideration in favour of strengthening the signalling capacities of states.

In addition to having the capacity for signalling, the state also should react to signals by other agents. If the USA signals a tendency to adopt a pro-cooperative preference and receives signals back that other states are also undergoing a similar preference change, the USA should then be more likely to settle on the cooperative preferences they tried out. The considerations of the Negotiated Order approach encourage changes which support such responsiveness. Again, the exact steps to take depend on the concrete situation.

The Negotiated Order approach points to changes that can encourage a positive result. We find ways to support a cooperative solution between states that did not become apparent from the perspective of standard rational choice theory:

- Render the identification of a problem more likely.
- Ensure that the group agent's structure is flexible enough to allow for preference change.
- Ensure preference change has the right direction.
- Enable the group agent to signal about preference change.
- Ensure the group agent reacts appropriately to such signals.

So far, we considered how changes to the agents help. Changing the context in which agents negotiate can also contribute to achieving a meaningful climate agreement.

Proposals for Changing the Context

This time we want the context to make it more likely that each agent

1. undergoes a transformation of its pay-offs, in particular,
2. introduces an i such that i > (T-R) and i > (P-S) during the exploratory phase, and
3. signals the potential preference change and settles on it if other agents signal a similar tendency.

The requirements remain the same as before, but this time we look at how the context can contribute towards meeting them. The international system and the negotiations serve here as the context in question. How can we change the context to increase the probability that the negotiating states will undergo a preference change in favour of climate saving cooperation?

First, the context should allow and incentivise a transformation of the game by the agent. In the case of climate change negotiations, we can push for an international system that enables states to change their course of action in a face-saving way, especially if the change is in favour of reducing climate change. The agreements between states should allow for amendments if a party becomes more cooperative regarding the climate negotiations.

I pointed out above that we face a trade-off between flexibility and credible commitments. Agents should be flexible enough to change their preferences, but also able to credibly commit themselves. We face an analogous trade-off on the level of the negotiation context. It must allow agents to change their preferences and act accordingly, and at the same time provide a background for stable agreements. The Negotiated Order approach, as developed thus far, does not show us how to balance such a trade-off successfully, but it draws our attention to the importance of such features for a positive scenario.

The context can also encourage the agent to develop an i such that i > (T-R) and i > (P-S) in a variety of ways depending on the exact mechanism of preference change. On my reconstruction of the Negotiated Order approach, agents enter an exploratory phase during which they try out tentative preferences. The feedback they receive influences which motivational profile they settle on and thereby can increase the chance of a sufficiently large i. Signalling comes in at this point, so that requirements (1) and (2) are closely related.

One salient option to increase the likelihood that other agents develop a preference for cooperation is to signal approval of a motivational profile that includes such a preference. Assume that India signals an exploration of becoming more cooperative. The EU might signal back that it approves of this potential change of preferences and might be willing to take steps to support India.[95] This option assumes that a signal of approval affects the chance of a motivational profile settling in positively. Whether this holds true must be judged in the concrete case. If Pakistan signalled approval, this might have a smaller or even the reverse effect.

The context must allow room for signalling, so that agents align in their intrinsic preferences. Accordingly, we should allow negotiations to take longer. By contrast, consider the Boadway et al. mechanism. In their model, group agents commit first to matching rates and second to direct abatement levels. Then the negotiations end.

Afterwards, the agents might still engage in emission quota trading, but overall abatement commitments remain fixed. Signalling has little place in this mechanism, since it is almost completely limited to declaring the matching rate. If one state picks high

[95] The EU's willingness in this scenario might be the result of prior preferences or of intrinsic preference change. We are concerned here with what advice we could give the agents, for example the EU, if they ask us how to increase the likelihood of forming a successful coalition of co-operators.

matching rates, this increases the subjective probability that it will also pick high direct abatement levels. But there is hardly any place for coordinating preference change.

The Negotiated Order model provides a reason to have multiple rounds of negotiations rather than to come to a quick Pareto-efficient agreement. We want the states to be able to undergo aligned pro-cooperative preference change to increase the climate friendliness of the resulting agreement. Again, we face difficult trade-offs. The climate is changing now. Adding another round of negotiations, while giving opportunity for signalling, costs valuable time. But the Negotiated Order approach at least gives us a reason not to endorse a quick and supposedly efficient mechanism such as the one proposed by Boadway et al. In sum, we find the following broad suggestions:

- The context should allow and incentivise appropriate transformations of the game.
- The context should leave room for signalling about preference change.

The Negotiated Order approach offers a suggestion on how to work towards successful climate negotiations, but how does it compare with the suggestion of standard rational choice theory?

Comparing the Negotiated Order Proposals with Boadway et al.

Before I turned to the Negotiated Order approach, I showed that standard rational choice theory offers guidance for how to structure negotiations to avoid unfavourable outcomes. The implementation theory model proposed by Boadway et al. suggests a mechanism for climate negotiations improving upon a prisoner's dilemma type scenario.

This implementation theory model has important advantages over the proposals I made from the perspective of the Negotiated Order approach: it proves, under certain assumption, that its outcomes are Pareto-efficient and therefore preferable to the prior

situation of mutual defection. It provides clear prescriptions: follow these two steps and the result has the following features (under certain assumptions).

In contrast, the proposals I just listed remained general and often came with a proviso such as that one must consider certain trade-offs or gather further empirical data. I offered no proofs but only plausible considerations in favour of certain changes. Given these shortcomings of my reconstructed Negotiated Order approach over standard rational choice theory, one might be tempted to stick to the latter. But that conclusion neglects the shortcomings of the model by Boadway et al. and similar models.

As I discussed earlier, the Boadway et al. two-stage negotiation mechanism only ensures a limited notion of optimality – Pareto-optimality. By introducing emission quota trading we can also equalise the marginal benefits of emissions over all states, but all these achievements fall short of moral requirements. Particularly if the states fail to take future generations into account, the Boadway et al. mechanism cannot deliver a positive result. Efficiency relative to the current preferences of negotiating states does not suffice.

We do not only want agents to follow their preferences Pareto-efficiently so that individual and collective rationality (in Gardiner's sense) coincide; in addition, we want states to have appropriate preferences. Boadway et al.'s model assumes fixed preferences and then shows how to achieve efficiency. The Negotiated Order proposals, on the other hand, provide a way of meeting more ambitious goals. The suggestions I offered remained general and in need of further empirical input, but they serve as a start towards achieving deeper cooperation through preference change.

Furthermore, the occurrence of preference change undermines the original success of the Boadway et al. model. A formerly efficient agreement on abatement levels is no longer efficient if the agents change their preferences afterwards. Even what the mechanism achieves becomes dubious if we look at the assumption of the model.

Both approaches, as they stand, have their shortcomings. The Negotiated Order proposals remain general and in need of further research, including empirical investigations and mathematical models. The implementation theory as represented by Boadway et al. provides a specific model, but one that rests on debatable assumptions and falls well short of our higher goals. Both offer valuable suggestions for ambitious climate agreements.

Future research can hope for a convergence of the approaches, for the Negotiated Order approach to spawn a quantitative formalisation, and for implementation theory to take intrinsic preference change into account. After incorporating future empirical research and improved models we might end up with specific recommendations that take preference change into account and allow us to achieve our ambitious goals regarding climate change. For example, the Boadway model was plagued by second-order worries, that is, by worries as to whether the states would even enter such a formal negotiation procedure. Perhaps the Negotiated Order approach and its accounts of preference change can help to overcome such difficulties which then allow us to apply models like that of Boadway.[96]

The Contribution

In this chapter, I applied the reconstructed Negotiated Order approach to group agents. I presented the differences the approach makes to the way we describe climate negotiations and the proposals we can derive from it. Although we found limitations in the approach in its current stage of development, it also made contributions going beyond

[96] I thank Yonatan Shemmer for this suggestion.

the standard rational choice model. The Negotiated Order approach serves as an original theory of negotiating group agents offering a unique perspective.

Climate change negotiations are a particularly important example. Humanity faces a threat in climate change and the Negotiated Order approach reveals different ways of tackling it. But the significance of the approach reaches beyond this one example. Negotiations between states and other group agents abound, and the Negotiated Order approach points to possibilities for negotiation that we do not see using standard rational choice models.

Beyond the illustrative example of climate negotiations, the Negotiated Order approach offers us a general theory of negotiating group agents. The journey has not reached its end, however, since research remains to be done. After the conceptual success of this thesis, there remains the work of formalising the models of the Negotiated Order approach and combining them with further empirical data.

Conclusion

The aim of the present thesis is to provide an account of group agents as negotiators who are potentially capable of solving the large-scale problems humanity faces, paradigmatically climate change. For this purpose, I have drawn on pragmatist social science and in particular the Negotiated Order approach. In chapters two to six we found a unified picture of motivational change in the interaction between agents including group agents.

The second to fourth chapter of this thesis aimed to provide a philosophically palatable reconstruction of the sociological Negotiated Order approach. The reconstruction included the development of a theory of preference change based on the account of motivational change found in the pragmatist literature. At the core of this theory was the idea of the Problematic Situation which opens agents up to a change of preferences. The notion of a Problematic Situation can already be found in John Dewey's work and served as the central theme of the second chapter of this thesis. Finding themselves disrupted in their course of action, agents can enter an exploratory phase during which they are more likely to undergo motivational change. Problematic Situations are supposed to have the force to open the agent up to a change of all types of preferences, including the intrinsic preferences over fundamental alternatives.

Chapter three combined the pragmatist account of Problematic Situations with the theory of signalling, which has become a hot topic in game theory over recent decades. Such an encounter of different approaches became salient because of the emphasis pragmatist sociology put on the symbolic aspects of human interaction. The resulting combination of exploratory preference change in response to Problematic Situations and signalling opened the door for new kinds of cooperation. Agents can signal to each other

what kind of preference change they might undergo. Having these elements in place, I offered a reconstruction of the Negotiated Order approach in the fourth chapter. While this reconstruction was selective, it allowed us to see how the Negotiated Order approach sketches dynamics which escape the view of standard rational choice theory.

Since the main interest of this thesis is in group agents as negotiators and problem solvers, my reconstruction of the Negotiated Order approach needed to be combined with a theory of group agency. The fifth chapter presents the two main accounts of group agency in the literature: functionalism and interpretivism. The chapter argues that interpretivism is seriously flawed and that our approach to negotiating group agents should endorse a version of functionalism. In addition, the fifth chapter provides constraints on what we should expect the realisers of functional states to be in group agents, given the Negotiated Order approach.

Chapter six serves as a proof-of-concept of the Negotiated Order approach applied to negotiating group agents addressing a real-world problem. It discusses current game theoretic models of climate change negotiation used in philosophy at length to then show the difference made by the introduction of preference change in response to Problematic Situations. The Negotiated Order approach allows us to see potential means of egress from dismal scenarios resembling a prisoner's dilemma or a tragedy of the commons.

In sum, the present thesis introduces a school of social science into philosophical discussions; in the course of doing so it puts forward a theory of preference change, and combines these ideas with a theory of group agency to allow a new perspective on how humanity might face large-scale problems. The final chapter shows that this proposal promises advantages over current standard game theoretical approaches. While the modelling of climate change negotiation certainly must go beyond what could be

sketched in this illustrative chapter, it indicates the benefits we can expect from following this path further.

Without doubt great challenges lie ahead of the proposed account of negotiating group agents. The present thesis does not offer a full formalisation of the theory of preference change (but see the appendix, which takes steps in this direction). Future research in how to combine the new picture of agency and motivational change with formal accounts in decision and game theory remains open. Nonetheless, through a large variety of topics the present thesis has managed to develop a unified theory of group agents as negotiators, which makes it possible to see how they could come to successfully address the large-scale problems facing humanity.

Appendix: Dewey's Decision Theory

The main part of this thesis has discussed the Negotiated Order approach and has drawn on Dewey's theory of the Problematic Situation for this purpose. But while this aspect of Dewey's work has exerted a clear influence on pragmatist sociology, it does not constitute the entirety of his thought on motivational change. Since it is not clear that these further elements have had as great an impact on the Negotiated Order approach, I have left them out of my reconstruction. Nonetheless, they provide an important background to pragmatist sociology and raise the question as to how to integrate the pragmatist tradition with decision theory. The purpose of this appendix is to provide a general take on Dewey's theory of practical reasoning and to underline the importance of preference change for the theory.

Apart from certain circles of committed admirers, Dewey's theory of practical reasoning[97] suffers from a lack of interest as a result of two widespread assumptions: First, Dewey's approach is incompatible with decision theory. Second, his proposed replacement and in particular his claims about ends and means make little sense, if any.

To overthrow both assumptions, I show that we can render Dewey's approach formal with a revised decision theory, and that the result allows us to make sense of Dewey's philosophy of action. His claims about ends and means become interesting and even plausible, once reformulated as contributions to decision theory.

Undoubtedly, Dewey discouraged attempts to quantify human decision processes.[98] According to him, we should resist the temptation to understand significant human

[97] In the present appendix I am especially drawing on Dewey's middle and late work, in particular his books *Human Nature and Conduct* and *Theory of Valuation*. Although there are differences in detail between Dewey's various formulations, I will assume an underlying unified theory.

[98] Other pragmatists have shown greater affinity to decision theory. Frank P. Ramsey (1926) was greatly influenced by pragmatism and his contributions to decision theory were revolutionary. For recent work on Ramsey's pragmatism, see Misak (2016) and Gruber (2017).

decisions by means of analogy with a stock investor trying to make the best deal (cf. MW 14: 151).[99] Despite Dewey's reservations, I will present three of his claims about ends and means and reconstruct them as contributions to decision theory:

First, new ends can arise out of impeded habits. If the environment blocks the realisation of habits, agents might change their ends in response. They start to explore different courses of conduct.

Second, ends are relative to situations. Dewey prefers to talk of ends-in-view, because for him ends only persist if they fulfil a certain function.

Third, means affect ends. Dewey claims that means and ends form a continuum rather than remain separate from each other. In a nutshell, information about how an agent can realise its ends alters these ends.

After introducing these three claims, I will show how we can reconstruct them as contributions to decision theory given a technical innovation: commitment values. These values range over intrinsic preferences and specify the probability that the preference will change by the next act. Dewey's claims give us rules for the dynamics of commitment values, that is, rules for how preferences change. I conclude after addressing concerns about the descriptive value of the proposed theory of preference change.

Before getting into the matter, a further complication deserves attention: Dewey disdained the separation of descriptive and normative inquiries. Accordingly, he would have rejected the distinction between descriptive and normative decision theory. While the discussion in the main text of this thesis remained descriptive and tried to provide an adequate account of how agents in fact undergo motivational change, this appendix

[99] In the quoted passage, he seems to have transformative decisions in mind, such as discussed by L. A. Paul (2016). See also Fesmire (2003: 76) for a discussion of this criticism.

follows Dewey insofar as it should have plausibility on a normative as well as a descriptive construal.

Dewey's Central Claims

Dewey criticises a strict application of means-ends schemes to human action and revises the notion of ends. All ends should be open to change, and they should resemble means more than traditional theories of practical reasoning assume.

While I do not reconstruct Dewey's claims as contributions to decision theory until later, I take the first steps in this direction here by translating the terminology from "means" and "ends" to "preferences" and "consequences". Although there is no trivial reduction of the folk-psychological to the decision theoretic vocabulary, I suggest that ends can be specified in terms of the preferences over consequences: to have an end is, other things being equal, to prefer consequences that realise it over other consequences. If my end is to read a book, I prefer consequences that realise this end over other consequences, other things being equal.

Ends Arising from Impeded Habits

As has been recognised in the secondary literature, Dewey's theory gives prominence to an idiosyncratic notion of habit. For Dewey, habits are dispositions that give action structure (cf. MW 14: 31, Anderson 2014). Without habits we would be driven by a bundle of raw impulses.

While Dewey's habits merit their own discussion, for the issue of motivational change the most important aspect of habits is their disruption. Sometimes habits become impeded

when the environment is contrary to the agent's expectations.[100] In response to such situations, agents exhibit exploratory behaviour and open up to motivational change. In effect, Dewey describes a special case of Problematic Situations brought about by the impediment of habits. Consider the following example of a Problematic Situation:

Matilda visits an alpine village every summer. She has a favourite path which she walks each year. On her current visit, she chooses to hike on this path, as is her habit.[101] Walking the path, however, she finds herself confronted with an unexpected obstacle. A tree has fallen across the path. Its size does not allow Matilda to climb over it. Thick thorn bushes on both sides of the path render even a small deviation from it difficult. The tree blocks the path and Matilda's activity.

Contrary to Matilda's expectations, her habit cannot give her activity structure. She responds with a phase of exploration with respect to the situation (cf. MW 14: 139-141). The exploratory process can take place in the agent's mind. Matilda goes through various options using her imagination. But she can also try out an end by acting as if she fully endorsed it. Finding her habitual route blocked, Matilda takes a few steps in this or that direction, looking down various paths before endorsing a final decision. She is opened up by the situation and explores the opportunities it offers. All of this should be familiar from the introduction of Problematic Situations in chapter two above, except that Dewey puts greater emphasis on habits.[102] In line with this previous discussion, such events of habit impediment and exploration alter the motivational life of agents:

[100] In the following, I will not always note that the impediment is unexpected but this is a silent assumption.
[101] Dewey's idiosyncratic notion of habit might also allow for one-time habits (cf. MW 14: 32). But this feature of his account doesn't make a difference for my purposes.
[102] My presentation here also does not mention the identification of problems, which I introduced as a necessary condition for Problematic Situations in chapter two. I assume silently that such an identification occurs so that these situations of habit impediment meet the necessary and jointly sufficient conditions for Problematic Situations.

> "Some habit impeded by circumstances is the source of the projection of the end." (MW 14: 29)

To project an end is to consciously endorse it. To project a new end entails being motivated by it. To project an old end is to reaffirm it. The situation of impeded habits opens agents up to such an endorsement of ends. Accordingly, to be opened up by a situation entails a propensity for motivational change.

But impeded habits do not guarantee that ends change. Matilda might double down on wanting to walk this path and project the old end. She might go get a chainsaw, call the fire brigade to remove the tree, or simply postpone her walk without giving up her end. Dewey endorses a strictly probabilistic theory of changing ends, according to which impeded habits render the change of motivations, such as Matilda picking a new favourite path, more likely.

Dewey does not restrict this change of ends to instrumental aims. All ends can be opened up when circumstances impede habits. If a change of non-instrumental ends occurs and Matilda develops the end of walking another path, we expect her to take this other path in the following year, even though she has no reason to assume that her formerly habitual path is still impassable, since non-instrumental ends do not depend on information. Accordingly, a decision theory for reconstructing Dewey's claims must include a mechanism for the change of intrinsic preferences, that is, preferences independent of information.[103]

My reconstruction must also account for the exploratory element in Dewey's theory, which I have pointed to throughout my thesis. Agents do not immediately endorse new ends in response to impeded habits, but rather explore how to move on after such a

[103] During the discussion of formal decision theory, I elaborate the notion of intrinsic preferences. See also Binmore 2009: 5-6.

situation. This exploratory phase poses a major difficulty for any decision theoretic reconstruction of Dewey: The agents apparently act without fully endorsing their preferences. According to traditional decision theory, one either has a preference or not, but one does not hold it in a qualified, exploratory manner for a certain period.

That new ends can arise out of impeded habits and that agents explore ends in response to such situations are not Dewey's only challenging claims about ends, as we will see.

The Horizon of Ends

To distinguish his theory of practical reasoning from traditional approaches, Dewey prefers the term "end-in-view" rather than "end". Dewey rails against understanding ends as fixed finalities that agents aim for in all situations (cf. MW 14: 159). He attacks the concept of the highest good, according to which we have a goal overruling all other goals at all times. Instead ends-in-view "arise out of natural effects or consequences which in the beginning are hit upon, stumbled upon so far as any purpose is concerned" (MW 14: 155). Ends have their origin in situations of action and must prove themselves to persist in such situations. As we have seen, impeded habits offer an occasion to abandon old ends-in-view by bringing about Problematic Situations. But even without such impediments, ends are not guaranteed an infinite lifespan.[104] Ends have to fulfil a function or they disappear:

> "Ends are foreseen consequences which arise in the course of activity and which are employed to give activity added meaning and to direct its further course." (MW 14: 155)

[104] One might think of the time preference approach used in economics. But while there might be interesting connections between these approaches, time preferences should not be confused with Dewey's proposal.

Ends serve the function of giving our activities meaning.[105] It would be wrong, however, to make this function another end of the agent. Then we would have a highest good all over again: the good of meaningful activity. Instead we should understand the function as governing the lifespan of ends. Unless ends fulfil their function, the agent loses them over time. Ends come with a sell-by date, which is pushed back if they do well. The function is not a highest good, but a constraint on the persistence of ends.

Assume Matilda goes for a walk and endorses the end of getting to know the wild flowers along the path. She looks at the blossoms and the leaves. Using her smartphone, she learns their names. Matilda's end of getting to know the flowers adds meaning to the activity of walking. But this end might not persist forever. Maybe on the next walk the end does not give her activity more meaning. The longer it fails to give meaning, the likelier Matilda is to lose it. We can think of the end as decaying, unless it receives a renewal by fulfilling its function.

In the following I assume that this horizon applies only to particular ends in the foreground of our motivations: ends that typically come to mind when someone asks us what we are trying to achieve (cf. Stevenson 1962: 93). Background ends, such as avoiding tripping over our feet, do not have to repeatedly prove themselves. In most situations, it would be odd to call health an end-in-view. It is an end, but it remains in the background rather than in view. My discussion will focus on the foreground ends.

For decision theory, Dewey's horizon translates into the claim that preferences that sustain foreground ends will fall out of use over time, unless they renew themselves by

[105] See also: "A hypothetical possible solution, as an end-in-view, is used as a methodological means to direct further observations and experiments." (LW 13: 232, see also Pappas 2008: 263) Dewey also suggests that ends-in-view are "means of unification and liberation of present conflicting, confused habits and impulses" (MW 14: 158). One should not overemphasise the role of habits and impulses for the function of ends, however. This quote belongs to Dewey's *Human Nature and Conduct*, which stresses habits and impulses more than other texts.

fulfilling their function. Extrinsic preferences are uncontroversially relative to situations, because they depend on information that differs between situations: if Matilda prefers to take the left over the right path because it leads her to the top of the hill faster, this preference remains relative to the situation. In another situation, where the right path leads to the top more quickly, she would take this path. These extrinsic preferences depend on the information about the paths. For *intrinsic* preferences, however, a dependence on the situation is controversial.

Standard decision theory assumes stable intrinsic preferences that hold in all situations. As commonly conceived, intrinsic preferences don't come with an expiration date. If Matilda intrinsically prefers knowing the wild flowers, then one would expect her to have this preference in other situations as well.[106] Not so with Dewey, as according to him, we are more likely to change our ends with new situations (cf. MW 14: 160-161). My reconstruction of Dewey's decision theory must make sense of the idea that foreground preferences change unless they add meaning to our activity.

For Dewey, ends have a horizon − are ends-in-view − because they do not differ that much from means. Just as means have the function to help fulfil ends, ends-in-view have their function as well. But Dewey goes even further in connecting ends and means.

Means Affect Ends

Perhaps the best-known element of Dewey's theory of action is the continuum of ends-means (cf. LW 13: 226), that is: Dewey's denial of the complete separation of ends and means. We can distinguish a strong and a weak version of the ends-means continuum.

[106] This still allows for the possibility of the agent to refrain from acting on these preferences when other considerations override it. For example, Matilda might be in a hurry to get to her philosophy seminar and lack the time to study the flowers.

The strong version does away with the distinction almost entirely. In this interpretation, to call a state of affairs an "end" is only done to add emphasis to it for present consideration, not to give it a *categorically* different status in practical reasoning. The following quote suggests the strong version:

> "The 'end' is merely a series of acts viewed at a remote stage; and a means is merely the series viewed at an earlier one." (MW 14: 27)

According to this passage, the difference between ends and means results from how much time one takes into consideration. Reaching the top of the hill is the end of the walk, because it remains remote, while taking the next turn is a means, because we face it soon. This strong reading of Dewey's continuum of ends-means eliminates the distinction almost completely.

By contrast, the weak reading of the continuum allows a categorical difference between ends and means but asserts that they stand in a stronger connection than usually presumed. While ends and means play different roles in reasoning, these roles involve each other closely. In this weak version, agents might change their ends in light of the means needed to achieve them (cf. Stevenson 1962: 95). I endorse this reading because it fits Dewey's motivation for denying the separation between ends and means.

The absurd results of clinging to an end regardless of the needed means motivate Dewey to introduce his continuum (cf. LW 13: 226-229). To take an example from Dewey, it is absurd to roast pork by burning down the whole sty of pigs.[107] If we cared exclusively about the end of roast pork, then we should see no problem in roasting pork by burning down pig-sties. Dewey believes there is a problem and therefore proposes his continuum.

[107] The example goes back to a humorous essay by Charles Lamb (see LW 13: 226-227). For a free online version visit https://www.gutenberg.org/ebooks/43566 [18. 8. 2017].

Within limits, Dewey's claim raises no controversy. Assume that Matilda has the end of going for a walk. She chooses reaching the top of the hill as an intermediate aim for her walk. However, after choosing the hilltop as an intermediate aim, she learns that the roads leading to it are badly maintained and dangerous. Upon learning this, she decides to walk to the lake instead. In this case, information about a means, the state of the path, overturned the intermediate aim of reaching the top of the hill. There is a continuum, at least of the weak version, between the means and the intermediate aim. This case is unproblematic, because the final goal was to go for a walk.

Dewey, however, extends his continuum to *all* ends, not just such intermediate aims. Returning to his original example, we find that such a general claim leads to difficulties. According to Dewey, if I learn that I cannot get roast pork without burning down sties, I should not only refrain from choosing roast pork for breakfast; in addition, I should be likelier to lose my end of eating roast pork altogether. Just as Matilda abandoned her intermediate aim of reaching the top of the hill, I might lose the end of having roast pork.

But why should means affects the end rather than the particular choice? I might choose to forego roast pork this morning, but that does not imply that I lose my general end of having roast pork. If someone told me that I can roast pork in an oven, I would do it.

For decision theory, Dewey's postulate translates to the claim that information about how intrinsic preferences can be realised affects these preferences. The information that I have to burn down sties to get roast pork is supposed to affect my preference for pork over a vegetarian meal. While it raises no controversy that such information can influence whether I act on a preference, it remains unclear how the information could lead to a *change* of my intrinsic preferences. Traditional decision theory lacks the means to formalise Dewey's claim because it assumes stable intrinsic preferences. Therefore, my reconstruction must go beyond traditional decision theory to capture Dewey's claims.

In sum, three desiderata guide my reconstruction of Dewey's decision theory: First, the reconstruction must specify how impeded habits lead to exploration and preference change. Second, the reconstruction must include the horizon of ends and the horizon-deferring effect of an end fulfilling its function. Third, the reconstruction must make plausible that information about how intrinsic preferences can be realised might alter these preferences. I now turn to decision theory and the technical innovation of commitment values to meet these desiderata.

The Formal Apparatus

To reconstruct Dewey's claims as contributions to decision theory I need a formal apparatus. This apparatus can be split into standard decision theory and the mechanism of preference change governed by commitment and described by commitment values.

Introducing Decision Theory

I start with a simple Savage-type decision theory. We have a set X of consequences.[108] For the sake of simplicity, I assume that these consequences are mutually exclusive and jointly exhaustive. For example, Matilda might either walk in the sun or walk in the rain or walk in the shade under trees, but not all or none of these options.

Preference relations hold over such consequences. For my purposes, strong preference and indifference are going to be the only preference relations.[109] The "$\succ$" sign indicates a strong preference. Matilda has a strong preference for the consequence of a walk in

[108] Hansson's (1995) formalisation of preference change also allows preference change by introducing or removing alternatives.
[109] In the literature, the "$\succcurlyeq$" indicates a weak preference, which is the disjunction between a strong preference and indifference relation. However, for expositional reasons that will become clear later, I avoid the weak preference relation.

the sun over a walk in the rain: *walk in the sun* $\succ$ *walk in the rain*. The "$\sim$" indicates indifference. Matilda is indifferent towards a choice between a walk under trees if the sun is shining and a walk under trees if it is raining, because neither the sun's rays nor rain can reach her there: *walk under trees during sunshine* $\sim$ *walk under trees during rain*.

Assuming that the preference relations over consequences are transitive and complete, we can construct a utility function.[110] The transitivity assumption states that if the agent prefers (or is indifferent between) consequences A to B and B to C, she also prefers (or is indifferent between) A to C. The completeness assumption states that for each two consequences in *X*, the agent either strongly prefers one over the other or the agent is indifferent. For all consequence pairs we find one preference relation.

Given these assumptions, we can assign utility values to consequences. For example, Matilda might derive a utility 10 from the consequence of a walk in the sun, a utility 0 from a walk in the rain, and a utility 5 from a walk under trees that block rain and sun:

$$u(\textit{walk in the sun}) = 10,\ u(\textit{walk in the rain}) = 0 \text{ and } u(\textit{walk under the trees}) = 5.$$

Agents have a set of available acts *A*, such as taking a path in the open or a path under trees. To model choice under risk, we assume that consequences are combinations of a state of the world, which has a certain subjective probability, and an action. For example, the consequence of a walk in the rain is a combination of the act of walking in the open and the state of it raining.

Each act comes with an expected utility value, calculated by adding the utilities of the potential consequences multiplied by the subjective probability of the state of the world. Consider Matilda's act of taking a path under the open sky, allowing rain and sun to reach her. Matilda's subjective probability for rain is high: $p(\textit{rain}) = .8$ and the probability of

[110] However, Mandler (2001) has argued that we should expect preferences only to fulfil one or the other but not both assumptions.

sunshine is low: $1-p(rain) = .2$. The expected utility of this act is: $eu(taking\ open\ path) = 0 * .8 + 10 * .2 = 2$. Matilda is better off walking under the trees: $eu(walking\ under\ trees) = 5 * .8 + 5 * .2 = 5$.

The subjective probabilities make no difference to the preference ordering over consequences. Matilda prefers a walk in the sun over a walk in the rain independently of how likely she considers sunny weather to be. But we can construct a second preference ordering over acts using expected utilities. According to our calculations, Matilda prefers the act of walking under trees over the act of walking in the open. In contrast to the preferences over consequences, this preference ordering over acts depends on subjective probabilities. If Matilda had given $p(rain) = 0$, then she would have preferred the act of walking in the open.

I call the preferences over consequences "intrinsic preferences" and the preferences over acts "extrinsic preferences", because the first are independent of subjective probabilities while the latter depend on them (cf. Binmore 2009: 5-6). Matilda intrinsically prefers a walk in the sun over a walk under trees. However, she extrinsically prefers the act of walking under trees over walking in the open, as our calculation has indicated.

Those are the outlines of the standard decision theory I assume. The theory alone cannot make sense of Dewey's claims, because it does not include any change of intrinsic preferences. The preference relations over the consequences are static in the picture we have inherited from Savage. For a successful reconstruction, we need more.

Introducing Commitment Values

For reconstructing Dewey's tenets as contributions to decision theory, the traditional approach lacks a way to model the change of intrinsic preferences. Such preference

change occurs if and only if a strong preference relation over consequences replaces an indifference relation or vice versa, or the relata of the strong preference relation switch places.[111] To model preference change, I introduce *commitments* and *commitment values*. According to my account, preferences are mental states governing choice behaviour, while commitments are properties governing the probability that these preferences change (see Strohmaier & Messerli typescript).[112] For example, if Matilda is highly committed to her preference for the green over the red party, but less committed to her preference for chocolate ice cream over vanilla ice cream. In other words, Matilda is likelier to undergo preference change with regard to ice cream than parties.

Commitment values represent the degree of commitments, that is, how likely preference change is. Commitment values are a further type of value in addition to utility values and subjective probabilities, and cannot be directly derived from these other values. Formally, commitment values range over the set of intrinsic preferences *P*, which contains all preference relations over consequences. For each intrinsic preference we have one commitment value. The commitment value of a preference specifies the probability that the preference changes before the next act begins, given that the agent's mind functions normally.[113] The normality clause rules out preference change by way of a brain seizure or a surgical intervention.

[111] As mentioned above, I ignore the weak preference relation. The reasons have now become clear: if we replace a strong preference relation with a weak preference relation, this could signify a preference change but it does not have to, because the weak preference relation is the disjunction between strong preference and indifference. Losing a preference relation without replacing it and establishing one in place of a previous gap would become additional types of preference change.

[112] I generally follow Dietrich & List's (2016) argument for mentalism about preferences. Those unhappy with postulating a preference as a mental state and commitment as a property of such states can accept commitment values but have to give them a different interpretation.

[113] The take of probability raises the question of whether we are concerned with objective or subjective probability. Both options are viable given appropriate background assumptions. If human minds are non-deterministic, the probability could be objective. If human minds are deterministic, I propose to interpret the probability as the subjective expectation of a preference change assigned by an idealised well-informed bystander.

In my proposal, the commitment value for an intrinsic preference is a real number between 0 and 1. That is, for all preferences p in *P*: $1 \geq com(p) \geq 0$. For example, Matilda has a high commitment to her preference *walk in the sun (WS) ≻ walk in the rain (WR)*, so that *com(WS≻WR) = .999*. Accordingly, there is only a .1 per cent chance that this preference will change by the next act. The chance that she either starts to prefer a walk in the rain or becomes indifferent in that time is small. If she had a commitment value of .5, then she would be as likely to change her preference as not.

I allow for a commitment value of 0 as a limiting case. A commitment value of 0 ensures that the agent will have changed preferences by the next act. Consider a quitting smoker who sincerely and with sufficient willpower acts one last time on her preference for a cigarette. We can describe her as having a preference at this point of time but with zero commitment to it. By the next act, she has lost the preference for a cigarette over no cigarette.

Commitment values range over all intrinsic preferences, but not over utilities or subjective probabilities. As it turns out, we can reconstruct all of Dewey's claims with such a limited set of commitment values. Nonetheless, the proposed decision theory shouldn't disconnect preferences from utilities. If Matilda loses her preference *WS≻WR*, a change in the utility values of *WS* and/or *WR* should reflect the preference change. The construction of utility function out of preferences standardly assumes transitivity and completeness. However, besides connecting utility values and preferences, the transitivity assumption entails restrictions for commitment values.[114]

Assume that Matilda has a preference for a walk in the sun over a walk in the rain *(WS≻WR)* and a preference for a walk in the rain over a walk during snowfall *(WR≻WN)*.

[114] Giving up the completeness demand, preference relations could also change by disappearing altogether rather than being replaced. In the following I assume this to be ruled out.

According to the transitivity assumption, Matilda also prefers a walk in the sun over a walk during snowfall *(WS≻WN)*. It follows that if Matilda underwent change regarding her preference *WS≻WN*, she must lose either her preference *WS≻WR* or *WR≻WN*. The probability of one preference change depends on the probability of the other preference changes.[115]

I proposed to interpret the commitment value of a preference as the probability of the preference to change by the time of the next act. From the interdependence of probabilities for preference change follows the interdependence of commitment values. They can only take values such that we are not forced to predict intransitive preferences.

We can formalise this constraint on commitment values. For illustration, I stipulate that an agent faces a set of three consequences such that $X = \{c_1, c_2, c_3\}$ and has strong preferences over these consequences such that $c_1 \succ c_2$, $c_2 \succ c_3$, and $c_1 \succ c_3$. The transitivity requirement mandates that given the first two, the last of these three preference relations must hold. Accordingly, the commitment values for the first two preference relations provide a floor for the commitment value of the third relation. The probability of a change of the last preference $(1- com(c_1 \succ c_3))$ cannot be larger than the combined probability of a change in the first two preferences:

$$1 - com(c_1 \succ c_3) \leq 1 - com\,(c_1 \succ c_2) + 1 - com(c_2 \succ c_3)$$

With the introduction of this transitivity restriction we have the tools at hand for reconsidering Dewey's claims. In my reconstruction, Dewey's theory of practical reasoning concerns the question of what makes a commitment value increase or decrease. As I will show in the following, Dewey describes and prescribes the dynamics of preference change.

[115] The same holds for indifference relations.

Reconstructing Dewey's Claims

Having introduced these commitments, I can reconstruct Dewey's claims as contributions to decision theory. Commitment values serve as a tool to describe Dewey's kinematics of ends.

Impeded Habits Decrease Commitment

Dewey proposes that the impediment of habits makes motivational change likelier. If habits cannot guide activity in virtue of unexpected obstacles, the commitment to associated preferences sinks. Finding her path blocked, Matilda's commitment to her action-guiding preferences decreases and, accordingly, the probability of preference change increases.

Before running into the tree, Matilda might have preferred the consequence that she walks path A over path B with a commitment value of .99. In light of the impeded habit, Matilda's *com* $(A \succ B)$ sinks to .4. There is a 60 per cent chance that she will have abandoned the original preference by the next act.

In a nutshell, the impediment of habits leads to an experience which reduces commitment and thereby affects the change of preferences. Giving experience such a prominent role fits well with Dewey's philosophy of action (cf. MW 14: 47, 133, see also Godfrey-Smith 2014). The more the agent experiences the impediment of her habits as problematic, the more the commitment value decreases. In the example, the more Matilda experiences it as problematic that the tree blocks her path, the lower her commitment to the preference for following this path.

I introduce a variable for describing the effects of such an experience: ex.[116] For reasons that will become clear later, the experience variable can also take a positive value so that it ranges between -1 and 1: $1 \geq ex \geq -1$, but for the special case of impeded habits the variable takes a negative value: $0 > ex \geq -1$. We calculate the new commitment value by adding the experience value for the prior situation to the prior commitment value:

$$com\,(A \succ B)_{t+1} = com\,(A \succ B)_{t} + ex_t$$

Of course, the general floor of 0 and ceiling of 1 for commitment values must remain intact, a limitation I silently assume in the following.

This formula presupposes that the experience values have ratio-scale commensurable with commitment values. Otherwise it would make no sense that a negative experience value leads to a reduction of commitment. While these are substantive assumptions, they offer the best interpretation of Dewey's decision theory.[117]

Which commitments are affected by an impediment? Presumably Matilda's commitment to her preference for chocolate over vanilla ice cream does not decrease because the environment impedes her habit to walk a certain path. This particular impediment should affect the particular commitments associated with the walk.

One might suggest limiting the commitment reduction to the preference which underlies the impeded habit, in Matilda's case the preference for path A over path B.[118] However,

[116] Dewey has a rich notion of experience. The experience variable introduced here is only supposed to capture one aspect of it, namely the extent to which the situation is experienced as problematic.

[117] For support of this reading see also Elizabeth Anderson's discussion of how experiences of trying out new valuations affect future valuations. "As the individual engages this new valuation, she experiences the consequences of acting on it. Reflection upon these consequences is then incorporated into more intelligent valuations, by way of further appraisals" (Anderson 2014). Anderson also reads Dewey as assuming a commensurability between experiences and practical attitudes.

[118] One might question whether each habit has an underlying preference. Maybe agents also have mere habits without underlying preferences. I bracket this question for the purpose of this reconstruction and assume there is always such a preference.

we should expect a broader effect, because the transitivity assumption renders commitment values interdependent, as I argued. Trying to keep the preference change minimal, we can take the commitment to preference for path A over B as the starting point of change, since it is directly associated with the impeded habit, and then calculate other changes using the equation for the transitivity restriction given above.

I have sketched how an impeded habit can lead to the projection of different ends: It decreases commitments and thereby affects the change of preferences as described by Dewey. So far, however, I haven't captured the exploratory phase. The change of ends in response to an impeded habit does not settle in immediately, instead the agent explores different courses of action. We have to formalise that during a stretch of time following the unexpected impediment of habits, agents relate in a special way to their preferences. Traditional decision theory lacks the tools to qualify the preferences for a period. Introducing commitments allows us to overcome that difficulty. After having her habits impeded, Matilda tries out various preferences with a low commitment until new preferences stick.

I propose the following approach to formalise the exploratory phase: We specify an interval from t+1 until t+n, the exploratory phase, during which the commitment to new preferences does not reach prior levels. This ceiling on commitment values for the relevant preferences characterises the interval. To be exact, during this phase no commitment value for a preference having A and B as relata can exceed the value of $com\,(A \succ B)_{t+1} = com\,(A \succ B)_{t} + ex_t$ where $0 > ex \geq -1$.

At the end of the exploratory phase the ceiling disappears. This implies neither that the commitment value after the exploratory phase is larger than during the phase, nor that the preferences at the end differ from those at the beginning. The phase's end only means that the commitment can take larger values again.

Specifying the length of this interval poses a major problem. It might depend on such factors as the nature of the original impediment and the experiences which follow during the exploratory phase. Dewey's empiricist tendencies limit how much we can abstract from the concrete situation. A pragmatist interested in decision theory has to confront the question of whether to follow Dewey here or simplify the rules for commitment values. Despite these concerns about how to further develop pragmatist decision theory, my reconstruction succeeds in formalising Dewey's claims and rendering them plausible. What seemed opposed to decision theory becomes a contribution to it, once commitment values have been introduced.

A Horizon for Preferences

Dewey prefers the term "end-in-view" over "end" because it indicates that ends come with a horizon. Ends-in-view only persist if they fulfil their function. My reconstruction captures this proposal by letting commitment values for a preference decrease over time. Matilda's commitment to her preference concerning wild flowers decreases, unless fulfilling its function re-enforces it.

We can conceive of the commitment value's reduction in analogy to a decay in nuclear physics. The decay function of carbon-14 describes that for any sample, only half the particles remain after a few years. Likewise, commitment values decay over time. The longer the time, the smaller the commitment value, the larger the probability of a preference change.

However helpful, the analogy remains imperfect. First, it is not clear that we should attribute an exponential decay as is typical for nuclear physics to commitment values. The decay-function might take a different form. Second, decay is not the only factor that affects commitment. The decay-function gives a baseline for the commitment to the

preference in the future, but other factors might change the commitment. We already encountered one of them: impeded habits lead to an additional decrease in commitment values.

How well the ends-in-view function is a further factor that affects the decrease of commitment. Dewey suggests that when ends-in-view function well, this might produce re-enforcing experiences reversing preference decay. If Matilda's end to inquire into the wild flowers functions well, then the experience re-enforces the commitment. Introducing such a re-enforcement allows reconstructing ends to have a function without postulating a highest end.

While Dewey does not spell out the details, the claim that positive experiences re-enforce the commitment suits his emphasis on experience. We can use the experience variable 'ex' introduced above. In effect, the equation for calculating the change of an intrinsic preference from t to t+1 should look as follows

$$com\,(A \succ B)_{t+1} = DF(com\,(A \succ B)_t) + ex_t$$

where *DF* is the decay function taking a commitment value at a time t and returning it reduced for t+1.

Specifying a non-arbitrary decay function poses a major problem. Should it be an exponential function as in nuclear physics? Then the function would take the form

$$\frac{d\,com\,(A \succ B)}{d\,t} = -\lambda\,com\,(A \succ B)$$

where λ is the exponential decay constant (ignoring the effects of experience). Even granting that the decay should be exponential, the question remains as to whether we have the same decay function for all kinds of preferences or whether the decay constant varies. Matilda's commitment to preferences concerning types of ice cream might decay faster than her commitment to voting preferences.

As I suggested in the beginning, we might want to distinguish foreground and background ends, where the latter do not exhibit a horizon. Some ends, such as the end to be healthy rather than to suffer from a sickness do not tend to disappear. Accordingly, some commitment values do not seem to decay at all.[119] The question as to which commitment values decay and why is left open by Dewey's texts and I too will leave it for future research.[120]

The questions about decay functions become even more difficult if we accept Dewey's theory of practical reasoning as a normative prescription. How should we adjudicate the rationality of various decay functions? Again, Dewey's text leaves the final answer open, but his general pragmatism suggests adaptation to the environment as a criterion for evaluating functions. The decay should not occur so quickly as to disrupt adaptation between the agent and the situation, but fast enough for re-adaptation. The challenge lies in cashing out the notion of adaptation so as to adjudicate between decay functions.

A further problem results from describing the decay as a function of commitment values over time, such that it takes a commitment value at a time and returns a diminished commitment value. This description cannot be completely correct, because if the decay function ranged over all moments of time, then commitment values would fail to correctly specify the probability of preference change by the next act.

As a solution, I let the decay function range over a quantised time. The decay happens from act to act rather than from moment to moment. Since the decay never occurs between acts, one can interpret commitment values as the probability of preference

[119] We could also capture this intuition by assuming countervailing factors reversing the decay, but it is hardly plausible that I have frequently positive experiences resulting from my preferences for health over sickness.

[120] Perhaps the best way to go here is to employ Dewey's notion of growth. The decay of some commitment values might contribute to growth while that of others might not. But Dewey's notion of growth is notoriously difficult to cash out.

change by the next act. One might even suggest a more coarse-grained time which only includes significant acts. Commitment to preference might only decay from one relevantly different situation of action to another. Consider a telephone operator engaging in repetitive acts, such as taking and forwarding calls. It could be that the decay progresses only when the operator switches from these tasks to other activities.[121] While I leave open which decay functions we should endorse, I succeeded in reconstructing Dewey's horizon for ends in a formal way by introducing commitment values.

Means Affect Commitment Values

Dewey's claim that means affect ends poses a challenge for every reconstruction. It is uncontroversial that ends can outweigh each other in case of a particular choice. According to Dewey, however, means affect the general ends and not just the occasional choices. I understand such general ends as specified by intrinsic preferences.[122] It seems implausible that any information could affect these preferences. Information is typically considered to influence subjective probabilities, not the preferences which are independent of such probabilities.

Consider the following example: Matilda has the end to walk to the top of the hill rather than to the lake. However, the only path leading there is badly maintained. At one point, Matilda would have to jump over a dangerous crevice. Taking such a path conflicts with her end of staying safe. Uncontroversially, Matilda's end of safety can outweigh her end of reaching the top of the hill.

[121] I thank Yonatan Shemmer for this suggestion and example.
[122] We can also take extrinsic preferences over acts to provide ends. However, only the first type of ends, which are given by intrinsic preferences, create a problem for Dewey's decision theory.

Dewey, however, makes the unusual proposal that learning about the cost of the means can lead to a change of the end itself. According to Dewey, Matilda might lose her end to get to the top altogether, so that she would not even go to the top of the hill if she learned of a different, safe path. Remember, if I learn about the need to burn down pig-sties for roast pork I should be likelier to lose my preference for roast pork. Even learning that I could use an oven instead, I would no longer do so.

Cast in these terms, we might wonder why Dewey lets the means affect the ends. Why should the mere information about how intrinsic preferences can be realised affect these preferences? Dewey's position appears in conflict with the most sensible interpretation of how information about means influences decisions. By introducing commitment values, however, we can render Dewey's position plausible. Instead of letting information about means change preferences directly, it affects the commitments governing the probability of preference change.

Upon receiving the information that realising the intrinsic preference comes with high costs, the commitment value should decrease. Matilda has an intrinsic preference for reaching the top of the hill and risking her safety over going to the lake. Accordingly, Matilda prefers to take the path and jump over the dangerous crevice. But she should be less committed to this preference for getting to the top, because it leads her to increase the probability of states she values negatively, such as incurring injuries. In light of the conflict with other ends, Matilda should consider whether the preference is worth it.

Dewey suggests that to unwaveringly stick with one's preferences if they overrule other ends repeatedly is irrational. An agent might be willing to risk her life for her cause. However, when it becomes clear that she can only further the cause by risking her life, it is rational for her to wonder whether she should keep the preference.

While an agent receiving the information that an intrinsic preference is likely to result in negative states should reduce her commitment, this does not rule out that some preferences have enough support to maintain a high commitment even though realising them precludes other positive states. Consider a resistance fighter living under a fascist regime who endures great pains for her struggle. She risks her life repeatedly for the just cause. We do not want to be forced by our decision theory to label her irrational and my reconstruction does not imply such an irrationality attribution. While the information about the means needed to fight the fascist regime should reduce the commitment value, conversely, other factors might raise the commitment value. The commitment reduction is ceteris paribus.

Using the technical innovation of commitment values, Dewey's assertion that we should not consider the ends separate from the means becomes plausible. We can also express it formally if we specify a measure of the costs for realising a preference. Such a measure would be easier to construct if we did not assume the exclusivity of consequences. Then we could calculate the cost of a consequence in terms of how many positive consequences it ruled out. But even assuming exclusivity, multiple ways to measure costs remain.

For example, when an agent receives information about the means needed to realise a preference, we can ask: How much would the agent have paid in US dollars before receiving the information in order for it not to be true? How much would Matilda have paid for the road not to include a dangerous crevice? We then calculate the new commitment value for the preference for the path to the hilltop over the other path (p) by deducting the amount of money (m) multiplied by a weighing factor (d):[123]

[123] I ignore the decay function and experience here.

$$com(p)_{t+1} = com(p)_t - m * d$$

There are problems with such a measure for costs. Arguably, the value of money diminishes the more an agent has of it. The sum of dollars might therefore fail to measure costs appropriately. We could try other approaches, for example we could ask how much pain the agent would have been willing to endure for the information to be false. The final measure does not matter much, as long as we can determine the cost of preferences and use it to reduce the commitment accordingly.

The general procedure has three steps: First, look at which currently chosen acts would no longer be chosen if not for this preference. Second, measure the costs of the action. Third, new information about these costs leads to a change in commitment.

As can be seen, commitment values allow us to capture Dewey's intuition that information about means affects ends in a variety of ways. The story about pig-sties and ovens appeared implausible, but the proposal that learning about the costs associated with a preference influences commitment to this preference is worth taking seriously. What seemed hard to accept becomes an interesting suggestion for decision theory in my reconstruction.

Problems

I have reconstructed Dewey's central tenets using an extended decision theory, which includes commitment values. Two problems, however, threaten the usefulness of the resulting decision theory.

Commitment Values Remain Underspecified

While Dewey's central tenets concern the dynamic of commitments, it remains underspecified. At multiple points in my reconstruction I was forced to say that only vaguely described factors influence commitment values. Experience played an important role, but I offered no way of measuring it. Neither did I settle on a specific decay function, nor on a way to calculate the costs which lead to a change of commitment values. Furthermore, I have offered no suggestions for how to determine the initial commitment value after a preference change.

The under-determination of commitment values results from the limited scope of the present text: I focus on reconstructing Dewey's theory of practical reasoning. What seemed completely contrary to decision theory has become a contribution to it by introducing commitment values. The remaining work needed for developing a more encompassing pragmatist decision theory does not lessen this achievement.

Many of the gaps could be closed by broadening the inquiry. Pragmatists do not have to stick to Dewey's word and could instead make original contributions. Formally specifying the full dynamic of commitments should stand on top of the priority list for developing a pragmatist decision theory. The gaps in the current theory follow from the limitations of Dewey's texts.

Too Much Preference Change

One might worry that introducing commitment values results in too much preference change. Assume that Matilda has a preference p for chocolate over vanilla ice cream such that $com(p) = .99$ remains constant over a long period of time. We should then expect that over 100 acts, Matilda's preference will change. In the course of a long day Matilda acts a 100 times, so that we should assume the occurrence of a preference change, even

though Matilda never encounters a situation where she had to choose between chocolate and vanilla ice cream. A possible response is to simply assume that the commitment values approach 1, but by adding Dewey's commitment decay to the picture it becomes clear that they won't stay there.

Frequent preference changes might seem implausible. Should we endorse a decision theory on which our preferences change many times a day? On a closer look, however, it becomes acceptable. That Matilda's preference concerning ice cream changes over the course of 100 acts does not imply that she sticks with her new preference. Matilda might simply be indifferent between chocolate and vanilla ice cream at one point during the day and switch back after a few acts. After all, there is no reason for the commitment value for the new indifference to be high. And if it is .5 at its origin, there is a 50 per cent chance that it is already gone again by the time of the next act. While preference change might be frequent, it can also go unnoticed without making a behavioural difference.

For some preferences, such as preferring health over sickness, we might propose that the commitment value typically approaches 1 and no decay occurs. But for many preferences frequent changes are plausible, especially if they tend to revert quickly and can go unnoticed. In fact, this account of preference change allows us to account for 'trembling hand' phenomena, where agents choose options which they otherwise disprefer. The trembling fits well with a commitment value close to but below 1, where the preference jumps quickly back after changes. The concern about too much preference change proves rather toothless, since a change of preferences is not that threatening if it has few or no consequences.

Conclusion

While Dewey avoided quantitative approaches, I have managed to reconstruct Dewey's claims about means and ends as contributions to decision theory. Impeded habits can lead to a change in ends, ends can be situational relative ends-in-view, and means can have a horizon. All these postulates suggest rules for an extended decision theory. Dewey's theory specifies the kinematics of commitments. It provides a model for preference change, and gives prominence to experience, the situational horizon of commitment, and information about how preferences can be realised.

My proposal shows that we do not have to insist on opposing Dewey's theory of practical reasoning to decision theory. Instead we can move forward towards a unified and sophisticated theory of human choice. Furthermore, it succeeds in recasting Dewey's apparently outrageous statements about ends and means as intriguing hypotheses about the dynamics of preferences. While we might not accept all of Dewey's suggestions, my reconstruction provides the basis for a pragmatist decision theory.

References

For *The Collected Works* edition of John Dewey's the following abbreviations are used, the numbers indicating the volume:

MW: Dewey, J. (1976). *The Middle Works, 1899–1924*. J. A. Boydston (ed.), Carbondale: Southern Illinois University Press.

LW: Dewey, J. (1981). *The Later Works, 1925–1953*. J. A. Boydston (ed.), Carbondale: Southern Illinois University Press.

Alger, J. M., & Alger, S. F. (1997). Beyond Mead: Symbolic Interaction between Humans and Felines. *Society & Animals*, *5*(1), 65–81. https://doi.org/10.1163/156853097X00222

Anderson, E. (2014). Dewey's Moral Philosophy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2014). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2014/entries/dewey-moral/

Axelrod, R., & Hamilton, W. D. (1981). The Evolution of Cooperation. *Science*, *211*(4489), 1390–1396.

Barwise, J., & Perry, J. (1981). Situations and Attitudes. *Journal of Philosophy*, *78*(11), 668–691.

Binmore, K. G. (2009). *Rational Decisions*. Princeton: Princeton University Press.

Block, N. (1978). Troubles with Functionalism. *Minnesota Studies in the Philosophy of Science*, *9*, 261–325.

Block, N. (1981). Psychologism and Behaviorism. *Philosophical Review*, *90*(1), 5–43.

Blumer, H. (1969). *Symbolic Interactionism: Perspective and Method*. Berkeley: University of California Press.

Blumer, H. (1977). Comment on Lewis' "The Classic American Pragmatists as Forerunners to Symbolic Interactionism." *The Sociological Quarterly*, *18*(2), 285–289. https://doi.org/10.1111/j.1533-8525.1977.tb01413.x

Boadway, R., Song, Z., & Tremblay, J.-F. (2011). The Efficiency of Voluntary Pollution Abatement when Countries Can Commit. *European Journal of Political Economy*, *27*(2), 352–368. https://doi.org/10.1016/j.ejpoleco.2010.10.003

Brennan, G. (2009). Climate Change: A Rational Choice Politics View. *Australian Journal of Agricultural and Resource Economics*, *53*(3), 309–326. https://doi.org/10.1111/j.1467-8489.2009.00457.x

Brooks, D. H. M. (1986). Group Minds. *Australasian Journal of Philosophy*, *64*(4), 456–470. https://doi.org/10.1080/00048408612342641

Bryant, M., & Stensaker, I. (2011). The Competing Roles of Middle Management: Negotiated Order in the Context of Change. *Journal of Change Management*, *11*(3), 353–373. https://doi.org/10.1080/14697017.2011.586951

Cohen, M. D., & Axelrod, R. (1984). Coping with Complexity: The Adaptive Value of Changing Unity. *American Economic Review*, *74*(1), 30.

DeCanio, S. J., & Fremstad, A. (2013). Game Theory and Climate Diplomacy. *Ecological Economics*, *85*, 177–187. https://doi.org/10.1016/j.ecolecon.2011.04.016

Dekel, E., Ely, J. C., & Yilankaya, O. (2007). Evolution of Preferences. *The Review of Economic Studies*, *74*(3), 685–704.

Dennett, D. C. (1987). *The Intentional Stance*. MIT Press.

Dennett, D. C. (1991). Real Patterns. *Journal of Philosophy*, *88*(1), 27–51.

Dietrich, F., & List, C. (2011). A Model of Non-Informational Preference Change. *Journal of Theoretical Politics*, *23*(2), 145–164.

Dietrich, F., & List, C. (2013). A Reason-Based Theory of Rational Choice. *Noûs*, *47*(1), 104–134. https://doi.org/10.1111/j.1468-0068.2011.00840.x

Dietrich, F., & List, C. (2016). Mentalism versus Behaviourism in Economics: A Philosophy-of-Science Perspective. *Economics and Philosophy, 32(2)*, 249–281.

Epstein, B. (2015). *The Ant Trap: Rebuilding the Foundations of the Social Sciences*. New York, NY: Oxford University Press.

Epstein, B. (2017). What Are Social Groups? Their Metaphysics and How to Classify Them. *Synthese*. https://doi.org/10.1007/s11229-017-1387-y

Fesmire, S. (2003). *John Dewey and Moral Imagination: Pragmatism in Ethics*. Bloomington, IN: Indiana University Press.

Fine, G. A. (1993). The Sad Demise, Mysterious Disappearance, and Glorious Triumph of Symbolic Interactionism. *Annual Review of Sociology*, *19*(1), 61–87. https://doi.org/10.1146/annurev.so.19.080193.000425

French, P. A. (1979). The Corporation as a Moral Person. *American Philosophical Quarterly*, *16*(3), 207–215.

Gardiner, S. M. (2003). The Pure Intergenerational Problem. *The Monist*, *86*(3), 481–500.

Gardiner, S. M. (2004). Ethics and Global Climate Change. *Ethics*, *114*(3), 555–600. https://doi.org/10.1086/382247

Gardiner, S. M. (2011). *A Perfect Moral Storm: The Ethical Tragedy of Climate Change*. New York: Oxford University Press.

Gilbert, M. (1992). *On Social Facts.* Princeton, NJ: Princeton Univ. Press.

Gilbert, M. (1993). Is an Agreement an Exchange of Promises?: *Journal of Philosophy*, *90*(12), 627–649. https://doi.org/10.2307/2940815

Godfrey-Smith, P. (2014). John Dewey's Experience and Nature. *Topoi, 33(1)*, 285–291. https://doi.org/10.1007/s11245-013-9214-7

Gruber, M. (2017). Ramsey's Pragmatic Approach to Truth and Belief: Ramsey's pragmatic theory of belief and truth. *Theoria*, *83(3)*, 225–248.

Grundig, F., Ward, H., & Zorick, E. R. (2001). Classical Theories of International Relations. In U. Luterbacher & D. F. Sprinz (Eds.), *International relations and global climate change* (pp. 154–181). Cambridge, MA: MIT Press.

Grüne-Yanoff, T., & Hansson, S. O. (Eds.). (2009). *Preference change: approaches from philosophy, economics and psychology*. Dordrecht; London: Springer.

Hansson, S. O. (1995). Changes in Preference. *Theory and Decision*, *38*(1), 1–28.

Hansson, S. O., & Grüne-Yanoff, T. (2012). Preferences. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2012). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/win2012/entries/preferences/

Hardin, G. (2009). The Tragedy of the Commons∗. *Journal of Natural Resources Policy Research*, *1*(3), 243–253. https://doi.org/10.1080/19390450903037302

Hook, S. (1959). John Dewey–Philosopher of Growth. *Journal of Philosophy*, *56*(26), 1010–1018.

Huebner, B. (2012). [Review of *Review of Group Agency: The Possibility, Design, and Status of Corporate Agents, List, Christian, Pettit, Philip*, by C. List & P. Pettit]. *Ethics*, *122*(3), 608–612. https://doi.org/10.1086/664942

Huebner, B. (2014). *Macrocognition: A Theory of Distributed Minds and Collective Intentionality*. Oxford University Press USA.

Joas, H. (1993). *Pragmatism and Social Theory*. Chicago: University of Chicago Press.

Joas, H. (1996). *The Creativity of Action*. (J. Gaines & P. Keast, Trans.). Cambridge: Polity Press.

Joas, H., & Knöbl, W. (2009). *Sozialtheorie: Zwanzig Einführende Vorlesungen*. Frankfurt am Main: Suhrkamp.

Lawford-Smith, H. (2015). What 'We'? *Journal of Social Ontology*, *1*(2). https://doi.org/10.1515/jso-2015-0008

Lewis, D. (1969). *Convention: A Philosophical Study*. New York: Blackwell.

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

Lewis, J. D. (1976). The Classic American Pragmatists as Forerunners to Symbolic Interactionism*. *Sociological Quarterly*, *17*(3), 347–359. https://doi.org/10.1111/j.1533-8525.1976.tb00988.x

Lewis, J. D. (1977). Reply to Blumer. *Sociological Quarterly*, *18*(2), 291–292. https://doi.org/10.1111/j.1533-8525.1977.tb01414.x

List, C., & Pettit, P. (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford; New York: Oxford University Press.

Ludwig, K. (2017). *From Plural to Institutional Agency: Collective Action II*. New York, NY: OUP Oxford.

Lycan, W. G. (1987). *Consciousness*. Cambridge, Mass: MIT Press.

Madani, K. (2013). Modeling International Climate Change Negotiations More Responsibly: Can Highly Simplified Game Theory Models Provide Reliable Policy Insights? *Ecological Economics*, *90*, 68–76. https://doi.org/10.1016/j.ecolecon.2013.02.011

Maines, D. R. (1982). In Search of Mesostructure: Studies in the Negotiated Order. *Urban Life*, *11*(3), 267–279. https://doi.org/10.1177/089124168201100301

Millgram, E. (1997). *Practical Induction*. Cambridge, Mass: Harvard University Press.

Millgram, E. (2001). *Varieties of Practical Reasoning*. MIT Press.

Mintoff, J. (2004). Is an Agreement an Exchange of Intentions? *Pacific Philosophical Quarterly*, *85*(1), 44–67.

Misak, C. (2016). *Cambridge Pragmatism: From Peirce and James to Ramsey and Wittgenstein*. New York, NY: Oxford University Press.

Mölder, B. (2010). *Mind Ascribed: An Elaboration and Defence of Interpretivism*. Amsterdam, The Netherlands ; Philadelphia: John Benjamins Pub. Co.

Nathan, M. L., & Mitroff, I. I. (1991). The Use of Negotiated Order Theory as a Tool for the Analysis and Development of an Interorganizational Field. *The Journal of Applied Behavioral Science*, *27*(2), 163–180. https://doi.org/10.1177/0021886391272002

Nozick, R. (1993). *The Nature of Rationality*. Princeton, N.J: Princeton University Press.

Ostrom, E. (2015). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge, United Kingdom: Cambridge University Press.

Pappas, G. (2008). *John Dewey's Ethics: Democracy as Experience*. Bloomington: American Philosophy.

Paul, L. A. (2016). *Transformative Experience* (Reprint edition). Oxford: Oxford University Press.

Peacocke, C. (1983). *Sense and Content: Experience, Thought, and Their Relations*. Oxford University Press.

Peirce, C. S. (1992). *The Essential Peirce: Selected Philosophical Writings*. (N. Houser & C. J. W. Kloesel, Eds.). Bloomington: Indiana University Press.

Pettit, P. (2014). Group Agents Are Not Expressive, Pragmatic or Theoretical Fictions. *Erkenntnis*, *79*(S9), 1641–1662. https://doi.org/10.1007/s10670-014-9633-x

Pettit, P., & Schweikard, D. (2006). Joint Actions and Group Agents. *Philosophy of the Social Sciences*, *36*(1), 18–39. https://doi.org/10.1177/0048393105284169

Preisendörfer, P. (2011). *Organisationssoziologie: Grundlagen, Theorien und Problemstellungen*. Wiesbaden: VS, Verlag für Sozialwissenschaften.

Ramsey, F. P. (1926). Truth and Probability. In: Braithwaite, R. B. (ed.), *The Foundations of Mathematics and Other Logical Essay* (pp. 156-198). McMaster University Archive for the History of Economic Thought.

Ritchie, K. (2013). What Are Groups? *Philosophical Studies*, *166*(2), 257–272. https://doi.org/10.1007/s11098-012-0030-5

Ritchie, K. (2015). The Metaphysics of Social Groups. *Philosophy Compass*, *10*(5), 310–321. https://doi.org/10.1111/phc3.12213

Rock, P. (1979). *The Roots of Symbolic Interactionism*. Palgrave Macmillan, London. https://doi.org/10.1007/978-1-349-04084-1_2

Rovane, C. (2014). Group Agency and Individualism. *Erkenntnis*, *79*(S9), 1663–1684. https://doi.org/10.1007/s10670-014-9634-9

Rowlands, I. H. (2001). Classical Theories of International Relations. In U. Luterbacher & D. F. Sprinz (Eds.), *International relations and global climate change* (pp. 43–65). Cambridge, MA: MIT Press.

Rupert, R. D. (2005). Minding One's Cognitive Systems: When Does a Group of Minds Constitute a Single Cognitive Unit? *Episteme*, *1*(3), 177–188.

Rupert, R. D. (2011). Empirical Arguments for Group Minds: A Critical Appraisal. *Philosophy Compass*, *6*(9), 630–639. https://doi.org/10.1111/j.1747-9991.2011.00420.x

Scott, S. (2009). Re-clothing the Emperor: The Swimming Pool as a Negotiated Order. *Symbolic Interaction*, *32*(2), 123–145. https://doi.org/10.1525/si.2009.32.2.123

Searle, J. (1990). Collective Intentions and Actions. In P. R. C. J. Morgan & M. Pollack (Eds.), *Intentions in Communication* (pp. 401–415). MIT Press.

Sheehy, P. (2006). Sharing Space: The Synchronic Identity of Social Groups. *Philosophy of the Social Sciences*, *36*(2), 131–148. https://doi.org/10.1177/0048393106287184

Skyrms, B. (1996). *Evolution of the Social Contract*. Cambridge; New York: Cambridge University Press.

Skyrms, B. (2004). *The Stag Hunt and the Evolution of Social Structure*. Cambridge, UK; New York: Cambridge University Press.

Skyrms, B. (2010). *Signals: Evolution, Learning, & Information*. Oxford; New York: Oxford University Press.

Smead, R., Sandler, R. L., Forber, P., & Basl, J. (2014). A Bargaining Game Analysis of International Climate Negotiations. *Nature Climate Change*, *4*(6), 442–445. https://doi.org/10.1038/nclimate2229

Snow, D. A. (2001). Extending and Broadening Blumer's Conceptualization of Symbolic Interactionism. *Symbolic Interaction*, *24*(3), 367–377. https://doi.org/10.1525/si.2001.24.3.367

Stevenson, C. L. (1962). Reflections on John Dewey's Ethics. *Proceedings of the Aristotelian Society*, 62(1), 77–98.

Stigler, G. J., & Becker, G. S. (1977). De Gustibus Non Est Disputandum. *The American Economic Review*, *67*(2), 76–90.

Strauss, A. L. (1988). *Negotiations: Varieties, Contexts, Processes, and Social Order*. San Francisco: Jossey-Bass.

Strauss, A. L. (1997). *Mirrors and Masks: The Search for Identity*. New Brunswick, N.J: Transaction Publishers.

Strauss, A. L. (2014). *Continual Permutations of Action*. New Brunswick, N.J: Aldine Transaction.

Strauss, A., Schatzman, L., Ehrlich, D., Butcher, R., & Sabshin, M. (1963). The Hospital and Its Negotiated Order. In E. Freidson (Ed.), *The Hospital in Modern Society* (pp. 147–169). New York: The Free Press of Glencoe.

Strohmaier, D., & Messerli, M. (Typescript). A Commitment-Based Theory of Preference Change.

Thomasson, A. L. (2016). The Ontology of Social Groups. *Synthese*, 1–17. https://doi.org/10.1007/s11229-016-1185-y

Tollefsen, D. (2006). From Extended Mind to Collective Mind. *Cognitive Systems Research*, *7*(2–3), 140–150. https://doi.org/10.1016/j.cogsys.2006.01.001

Tollefsen, D. (2015). *Groups as Agents*. Malden, MA: Polity.

Uzquiano, G. (2004). The Supreme Court and the Supreme Court Justices: A Metaphysical Puzzle. *Noûs*, *38*(1), 135–153.

Vanderschraaf, P., & Sillari, G. (2013). Common Knowledge. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2014). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2014/entries/common-knowledge/

Weber, M. (1990). *Wirtschaft und Gesellschaft: Grundriss der Verstehenden Soziologie*. (J. Winckelmann, Ed.). Tübingen: Mohr.

Wendt, A. (1994). Collective Identity Formation and the International State. *American Political Science Review*, *88*(02), 384–396. https://doi.org/10.2307/2944711

Wilson, R. A. (2001). Group-Level Cognition. *Philosophy of Science*, *3*(September), 262–273.

Wood, P. J. (2011). Climate Change and Game Theory. *Annals of the New York Academy of Sciences*, *1219*(1), 153–170. https://doi.org/10.1111/j.1749-6632.2010.05891.x

Acknowledgments:

My research has profited from discussions with Bob Stern, Holly Lawford-Smith, Yonatan Shemmer, Brian Epstein, Luca Barlassina, James Lewis, and various other people. Those who have contributed in more personal ways will receive my gratitude in person.

The research was funded by the British Arts & Humanities Research Council via the White Rose College for the Arts & Humanities (grant AH/L503848/1).