Political Equilibrium

Justified Social Order in a Diverse Society

Universität Hamburg
Fakultät für Wirtschafts- und Sozialwissenschaften

Dissertation

Zur Erlangung der Würde eines Doktors der
Wirtschafts- und Sozialwissenschaften
„Dr. phil.“
(gemäß der PromO vom 24. August 2010)

Vorgelegt von

Florian Wieczorek

aus Oldenburg

Hamburg, Juli 2021

Vorsitzende: Prof. Dr. Anke Gerber

Erstgutachter: Prof. Dr. Peter Niesen

Zweitgutachter: Prof. Dr. Thomas Schramme

Disputation: 8.7.2021

urn:nbn:de:gbv:18-ediss-94131
https://orcid.org/0000-0003-4856-6620

Abstract

This thesis is concerned with the question of justified social order in complex and diverse societies. The first step in giving an answer consists in the rejection of the answer typically given within the familiar Rawlsian paradigm: the construction of hypothetical agents, facing a hypothetical choice situation, from which normative principles of morality or justice are derived. The second step thus consists in providing an alternative methodological framework made up of three building blocks. One, a descriptive account of the object of normative theorizing. Two, a model of justified social order as an open-ended ideal and, three, a procedure and testing conception relating the ideal back to social reality. In a third step I spell out this approach in form of a theory of justified social order. The normative core of this theory consists in the idea of justified social order as a compromise with psychological ownership. That is to say that social order is at least a compromise to the mutual benefit of everybody governed by that order. Ideally, it is further endorsed by all individuals to the extent that they can actually identify with a given set of norms as *their own* order. In a fourth and final step I argue that this open-ended ideal can be related to social reality by thinking of a social mechanism of norm selection and by adding a testing conception of justified social order. I only begin to sketch the contours of both ideas. My reflections on mechanism design bring together the field of deliberative democracy and a range of other ideas relevant to reasonable political decision-making. The idea of a testing conception is pursued by thinking about a comparative index of justified social order, in analogy to indices of democracy. In a nutshell, this inquiry argues for a procedural answer to the question of justified social order while replacing several normative notions commonly found in justificatory liberalism by notions that have an empirical and especially psychological underpinning.

Zusammenfassung

Thema dieser Dissertation ist die Möglichkeit und Ausgestaltung gerechtfertigter sozialer Ordnung in komplexen und diversen Gesellschaften. Ein wohl bekannter Ansatz um diese Frage zu beantworten besteht darin, eine hypothetische Entscheidungssituation zwischen hypothetischen Akteuren theoretisch zu konstruieren und daraus normative Prinzipien der Moral oder der Gerechtigkeit abzuleiten. Diese vor allem durch John Rawls bekannte Vorgehensweise erweist sich jedoch bei genauerer Betrachtung als ungeeignet. Je nach Auslegung, so meine Kritik, scheitert der hypothetische Ansatz entweder an der Überwindung des Pluralismus oder bedient sich illegitimer Abstraktionen. Daher argumentiere ich in dieser Arbeit sowohl für einen alternativen methodologischen Ansatz als auch für eine alternative Theorie gerechtfertigter sozialer Ordnung. Der hier vertretene Ansatz besteht darin, theoretisch-normative Überlegungen systematisch auf ihren tatsächlichen Gegenstand zu beziehen – eine Gemeinschaft aus Individuen und ihre Normen. Konkret besteht mein methodologischer Vorschlag aus drei Bausteinen: erstens, einer deskriptiven Konzeption sozialer Ordnung, zweitens, eines offenen und prozeduralen Ideals, und drittens, einer Übersetzung des Ideals in die soziale Wirklichkeit. Meine Theorie gerechtfertigter Ordnung veranschaulicht diese Vorgehensweise. Konkret gründet sich die Theorie auf ein Verständnis von sozialer Ordnung als eine Menge von sozialen Normen – Regeln, welche unser Verhalten alltäglich koordinieren und so stabile und kooperative Gesellschaften ermöglichen. Der Idealzustand der hier vertretenen Theorie besteht darin, dass wir die gemeinsame soziale Ordnung als einen wertvollen Kompromiss befürworten und dadurch als unsere eigene Ordnung annehmen können. Eine solche Identifikation mit der gegebenen sozialen Ordnung wird in der politischen Theorie schon länger als erstrebenswerter Gegenpol zur Entfremdung diskutiert. Der weiterführende Beitrag dieser Arbeit besteht darin, diesen Idealzustand mit Hilfe eines psychologischen Konstrukts namens "psychological ownership" genauer zu erfassen. Das entscheidende normative Moment des angestrebten Idealzustandes besteht in der Befürwortung von und Identifikation mit sozialer Ordnung auf der Grundlage wohlüberlegter Gründe. Darüber hinaus zeigt die Theorie, dass sich ein solcher Idealzustand tatsächlich nur im Rahmen eines öffentlichen Diskurses anstreben lässt. Schließlich diskutiere ich wie das Ideal in realen Gesellschaften mittels eines politischen Mechanismus verfolgt werden kann und skizziere einen Index zur Überprüfung der Annäherung an das Ideal gerechtfertigter sozialer Ordnung.

Contents

Introduction

> *"I'd always rather work something out then read it in the published papers. It's just more fun. The trouble about reading the papers first; I mean you go to a new field you'd think the right thing to do would be to read what everybody else has said about it. If you do that, you finish up accepting their questions as well as their answers as being what matters in the subject. And maybe they are wrong. Maybe they are asking the wrong questions. If you go into it ignorant like I do and read as little as possible and plunge in and try to solve things; occasionally you make a complete idiot out of yourself. And I do that quite often. But I don't mind too much. People are used to it by now. But occasionally I ask a question that other people haven't asked and provide an answer that other people haven't provided. And you know, that's much more fun."*
>
> John Maynard Smith in an Interview to the BBC (1995)

Originally I was draw to studying normative social theory by a naive question: What is the good society? In the past, thinkers attended this question head on and "dreamed"[1] about ideal societies such as Plato's Polis or Morus' Utopia. From the perspective of contemporary normative theory, however, this seems an impossible question to answer because today's theorizing usually starts out with the presumption of persistent disagreement about the good everywhere. And even if there was some objectively true or correct optimal social state to be discovered in theory, so far we seem to lack the methodology to find it. Hence, contemporary normative theorizing has retreated to the discussion of abstract principles of justice, freedom, morality or equality, which are occasionally related to questions about how different institutions realize these principles.

[1] Gerald Gaus (2016: I.1.4) calls "dreaming" a kind of ideal theorizing that is not concerned with questions of attainability.

Guided especially by classical and contemporary social contract theory, I have settled on a more technical and modest question: What is justified social order? We will specify and clarify this question in *Chapter 1.* Essentially it is asking whether and how we can have a society of diverse individuals who disagree about the good, but nevertheless affirm a set of common rules that structure and coordinate their social lives.

In asking this question I follow the tradition of social contract and public reason theory. In this tradition we approach social questions of the good by pointing to what can be confirmed by each individual citizen. In classical social contract theory this is reflected in the presumption of an *original agreement* of all citizens. However, this idea is susceptible to the obvious objection of being unrealistic. Hence, public reason theory has replaced the idea of comprehensive agreement with the idea of comprehensive justification, *revealing the reasons* all citizens have in favor of their social order.

Although I take many notes from public reason theory, there are some variations of it that I reject. One such departure is my rejecting of what I call *hypothetical choice modeling.* This common approach in public reason theory translates the question of what reasons citizens have into the question of what idealized version citizens *would* have under counterfactual conditions specified in theory. In *Chapter 2* I offer a systematic criticism of this approach and ultimately argue that we need an alternative.

This alternative is developed in *Chapter 3* under the label "Embedded Constructivism". Embedded Constructivism does two things fundamentally different than typical accounts in public reason theory. First, it avoids substantial normative conceptions of justice and morality. These very common conceptions are, in my view, pseudo technical terms, leading primarily to never-ending controversies. In its stead, Embedded Constructivism places descriptive and empirical models of social order, which are subject to more productive debates. Second, Embedded Constructivism replaces hypothetical choice modeling with the construction of an open-ended ideal that can be pursued and thus spelled out by actual citizens.

In terms of the history of normative thought, my argument is that the move from actual to hypothetical agreement was a mistake. It is the idea of actual agreement, or rather how we can realistically think about its materialization, that gets us to interesting answers of practical relevance. A good analogy is perhaps democracy: We do not get to any interesting answers by imagining the policies that would result from government by the people under ideal, counterfactual conditions. To the contrary, the interesting answers emerge when we think about how and to what extent government by the people can actually be done by the people.

What my arguments of the first two chapters also illustrate is that in normative social theory it is almost impossible to tackle a substantial question without also tending to fundamental

methodological questions. This is because there is no paradigmatic and widely accepted understanding of how one ought to profess normative social theory. Hence, before even starting to answer my research question, I need to answer the methodological question of how the substantial question is to be approached. This is why I spill so much ink on rejecting hypothetical constructivism in public reason theory and argue for Embedded Constructivism.

In *Chapter 4* I finally turn to answering my research question by constructing an open-ended ideal of justified social order. The core idea emerging in the construction is that, while there is a continuous range of more or less justified orders, there are also important qualitative differences of how individuals and their social order relate. Simply put, the ideal scenario I propose is that individuals endorse social order as a compromise in light of what is important and valuable to them as the persons they are. This kind of personal endorsement can facilitate deep psychological attachments of *ownership* of citizens to *their* social order. This is highly desirable, because social order as a compromise with ownership is likely to give us an order we want, that works well and includes advanced forms of cooperation.

Chapter 5 takes up the challenge of starting to think about how the open-ended ideal could be pursued by actual citizens. In detail, I begin to sketch, fistly, a *democratic proceduralism* that points to how actual citizens could achieve the ideal of justified social order and, secondly, a test of whether and to what extent the ideal state has already been achieved. Overall this is the most incomplete chapter of the inquiry, which is, on the one hand, due to the complexity of the challenge and, on the other hand, due to the fact that here we reach the limits of how far Embedded Constructivism can be pursued in theory.

This endpoint may perhaps seem unsatisfying, but the hopeful message of this inquiry is that a complete and reasonable utopia may still be specified down to the last detail for most given societies. It simply requires a lot more work by a lot more people and not just the theorizing of one theorist. Enough for the preliminaries – let's get to work.

Chapter 1

Topic, Problem and Approach

It is always a good idea to start with a question. The question we are concerned with in this inquiry is a standard question of public reason liberalism: How can the social rules that structure and coordinate the actions within a given community be justified to each individual? In this first chapter we will not attempt to give an answer. Instead, the goal is to get a precise understanding of what the question is asking. To this end I mostly rely on simple models, which are meant to display the core concepts and problems of this inquiry, while leaving the details and complexities to be dealt with in later chapters.
We proceed by firstly clarifying the two basic concepts of this inquiry: *social order* and *justification*. Secondly, we turn to the explication of the core problem in respect to the justification of social order: justification in a complex and diverse society. Thirdly and finally, I propose an approach for tackling this problem, compressed into a guiding principle of justified social order.

1.1 Topic: Justifying Social Order

The topic of this inquiry is the justification of social order. Thereby 'justification'[1] indicates its normative nature: Ultimately, I am not interested in the descriptive question of how human societies are structured, but in the normative question of how they *should* be structured. Nevertheless, and this is a basic theme of this inquiry, it is of crucial importance to also have an answer to the descriptive question of what social order is and how it works. Without such an understanding of the object of normative theorizing, we lack orientation. Thus we may fail at constructing a normative theory that fits its non-normative object and is of any practical relevance to actual societies. As a consequence, a normative theory that is not oriented by a descriptive understanding of its non-normative object(s) is likely to appear as a mysterious

[1] I use single quotation marks in order to indicate when I am referring to the concept of something.

and useless conundrum to us as actual social beings living under some given social order.
To avoid this pitfall, we start out with a descriptive explication of the phenomena and concept of social order. More in depth literature on the emergence and working of social order is provided in the first sections of *Chapter 3* and *Chapter 4*.

1.1.1 Social Order

What is social order? Many things may come to mind here. Walking through our daily lives we see order and structure everywhere: We navigate through well-thought-out and highly regulated traffic systems. We cooperate with others at our place of work according to a complicated web of laws and contracts. We buy things at supermarkets by exchanging fiat money. In restaurants we display good manners, tip the waiter and on the news we hear about governments passing laws and constitutional courts making important rulings. It seems, no matter where we go, every social situation we encounter is somehow ordered.

The Fact of Social Order

Social order is not a philosophical invention but a fact of social life. It has therefore quite appropriately been called the *grammar*[2] or *cement*[3] of society. Thus I take social order to be a *fact* of social life: "[...] every society has or is a social order." (Baier 1995: 201)
Starting out with the fact of social order implies that there is no need for a justification of such orders per se. That is, I assume that human beings are social beings and that stable interactions and communal life of such beings requires some entrenched framework of coordination. It is of course true that human beings can and in particular cases have decided against social life and thereby against living under any form of social order. Also, phenomena such a secession, (civil) war and terrorism might show that, for some unfortunate combination of individuals, there is no order they strictly prefer to having no order at all. Such cases aside, having an order in place that coordinates our actions and allows us to engage in large scale cooperation is obviously extremely beneficial and does not require any further justification.
This not to deny that the radical anarchist question of whether having a social order per se can be justified, is a meaningful and interesting question. It can for instance be interesting to think about life without rules in order to understand the advantages and disadvantages of social order. Thomas Hobbes' state of nature thought experiment is a famous case in point. Furthermore, it can be interesting for individuals to ask themselves if they would prefer to live outside of the shackles and blessings that come with living in community. Finally, thinking past

[2]I am referring here to Cristina Bicchieri's *The Grammar of Society - The Nature and Dynamics of Social Norms* (2006).
[3]This metaphor stems for John Elster's *The Cement of Society - A Study of Social Order* (1989).

the fact of social order could reveal entirely new ways of human coexistence and cooperation that we might not have thought of so far.
This being said, I assume the fact of social order to be empirically and normatively uncontroversial: Human beings generally do *and* want to live in stable communities, which requires a common set of rules, coordinating expectations and regulating interactions. The normative question I am concerned with here is not if such rules are justified per se, but how we can say that a particular set of such rules is justified for the particular community it governs.

From Rules to Norms

So far, all we have is a very broad phenomenon. In order to narrow it down, let us focus on social rules. That is to say that I ask you to put aside thoughts about physical objects such as courtrooms, parliament buildings or dollar bills as well as thoughts about officials such as government clerks or police officers. These things may very well be highly relevant to a complete treatment of the phenomena of social order. But this inquiry is not about a complete account of social order. It is about a normative perspective on social order. In this context, we focus on social rules because the rules bring out the normative character of social order and this is what eventually triggers the question of justification.
What then are social rules? Social rules are general prescriptions to do or not to do some action X in some situation S. Where do such rules exist? We can observe social rules in the form of behavioral patterns in a given society. For instance that (almost) all cars stop at red traffic lights. If such behavioral patterns are based on a genuine social rule, this rule also exists in our minds in the form of expectations. That is, expectations about what other people will do and about what others expect that we should do. Social rules may further exist in our minds in the form of internalized rules that trigger emotional reactive attitudes – instant and often powerful signals about what should be done.
This last form of existence – i.e. rules as internalized rules – is neither necessary nor sufficient for the existence of a social rule. Necessary and jointly sufficient for the existence of a social rule is that it actually structures social behavior and that it is upheld by a web of symmetric expectations, shared among those involved in the respective social practice. Such rules are essentially what Christina Bicchieri (2006) calls "social norms". I will lay out her theory of norms in more detail at the beginning of *Chapter 3*. For now, I adopt her terminology of speaking about "norms" rather than "rules". This choice of words is motivated not only by Bicchieri's excellent theory of social norms, but also by the fact that speaking of "norms" focuses our attention on what matters about social rules from the perspective of this inquiry: their normative character.
The normative character of norms consist in the demands they make on us to do or not to do

something. When stopping at a red traffic light, when bringing a present to a birthday party, responding to a greeting, or tipping the waiter, we experience this normative force to do or not to do something. Depending on how we look at the situation and whether we have internalized the norm in question, this demand may appear to be coming from within ourselves or rather from the surrounding community. In any case, we can literally feel that it is there and it is this normative demand that may trigger a demand for justification. That is, if confronted with a demand to do or not to do something in a given situation, this may trigger an individual to wonder: "I sense that there is a general requirement to do this, but why should I follow it?"

A Simple Model

We will return to the matter of justification in a moment. For now let me summarize what has been said so far in a simple model of two sets: Social order is a set of social norms that governs the behavior of a set of individuals. I use 'social norms'[4] to indicate, firstly, that the rules we are talking about here structure the *interactions* of a set of individuals. This specification is important to put aside rules that do not regulate social interaction, such as rules that individuals set for themselves, or rules that do not effectively govern social behavior, such as rules that are being ignored. Secondly, social norms always have a normative character in that they make a general demand to do or not to do something. This specification distinguishes social norms from what is usually referred to as "conventions", i.e. useful rules that help us to coordinate our behavior but do not make a normative demand of compliance.

Given the simple model of a set of social norms and a set of individuals, there are at least two natural ways of adding more complexity to both sets. Here I briefly reject both of them as unnecessary.

One objection against my simplistic model might state that the very broad conception of social norms so far glosses over some important distinctions in the realm of norms. More specifically, one might demand at least a basic differentiation between formalized and non-formalized norms, e.g. between legal and "moral" norms. However, I consider the importance of this distinction to be generally exaggerated. A detailed argument in defense of this claim can be found in *Subsection 3.1.2*. In short, the reason for neglecting the distinction between formalized and non-formalized norms is that both types of norms basically function in the same way through a web of symmetric expectations and are equally relevant to a theory of justified social order.

Another demand for more complexity may arise in respect to the idea of a set of individuals. This set could be many things. It could be a small group of friends, members of a golf club, inhabitants of a town, citizens of a nation state or all human kind. You might be tempted to

[4]I use single quotation marks in order to indicate when I am referring to the concept of something.

insist that we should distinguish between these very different sets because they must imply different kinds of justifications. There is something to this demand to be more specific about the set of individuals we are concerned with. Thus I add some further assumptions in the two following sections of this chapter. Until we get to these points, there is a simple solution we can stick to while trying to get our heads around the topic of this inquiry.
This solution consists in the idea that the existence of a norm triggers the existence of a respective group addressed by this norm. In order to figure out who is a member of a relevant group we can simply look at the norm itself and how it applies in practice: If a given norm N applies to a certain set of individuals S, S constitutes a group relative to N. So instead of worrying too much about the nature of the group, we can think of the jurisdiction of the norm, which atomically triggers the existence of a group. Namely, all those individuals to whom the norm applies.[5] This way we can ignore the difficult questions of what exactly constitutes a 'group', 'community' or 'society'. I will continue to use these terms fairly loosely. The important bit for now is the image of a norm, making normative demands on a set of individuals and thus triggering the existence of a group that constitutes the jurisdiction of that norm.

1.1.2 Justifying Social Order

This brings us to the matter of justification. "The problem of justification arises when moral authority is claimed: it is our fellow participants in the rule governed practice on whom we make demands in its name." (G. Gaus 2011: 268) Essentially the justification of a norm is the task of giving reasons to an individual as to why she should adhere to the norm in question. So in terms of our simple model, this is to say that there is a reflexive relation between the set of norms and the set of individuals: The norms relate to a group of individuals in that they make normative demands on a range of individuals, and these individuals may relate back to the norms in that they may ask for justification of these normative demands. In this context justification establishes an argumentative link between a norm and the individual addressed by that norm.
In spite of the abstract talk of sets of norms and sets of individuals, justification is first and foremost a straight-forward linguistic practice we are all familiar with. More generally speaking, this practice starts with someone making a claim of some sort. This claim does not need to be normative. It could also be a simple claim about what is the case, such as the claim that most cars stop at red traffic lights. As with the normative claim that cars ought to stop at

[5]The idea of a norm triggering a group is proposed by Peter Niesen (2017). On a similar note Gerald Gaus states that "the range of the public is determined by the extent of the moral practice governed by the rule." (G. Gaus 2011: 268)

red traffic lights, such propositions imply validity claims which may be challenged. (Habermas 1984: 15-16) If they are, the claimant is asked to present appropriate reasons in defense of her claim. Giving such reasons is what we call a justification.[6]

In summary, justification is the practice of giving reasons to someone in favor of some challenged claim. In our case this claim is a normative demand implied by some norm. The someone is an individual confronted with such a demand, whereby the giver and receiver of the justificatory argument can be the same person. This happens when we reason about a normative demand only in our own minds. But perhaps the more interesting and typical case is a discussion between two or more individuals where one side is demanding and the other side is giving a justification.

A Normative Justification Principle

So far we have been mainly concerned with a descriptive account of social order and justification. Now, how do we get to a normative perspective? In public reason liberalism it is common place to ground a normative principle of public justification in some universal claim to freedom and equality, naturally possessed by each individual. Summarizing this core liberal view, Jonathan Quong writes:

> "Public reason requires that the moral or political rules that regulate our common life be, in some sense, justifiable or acceptable to all those persons over whom the rules purport to have authority. [...] Proponents of public reason often present the idea as an implication of a particular conception of persons as free and equal. Each of us is free in the sense of not being naturally subject to any other person's moral or political authority, and we are equally situated with respect to this freedom from the natural authority of others. How, then, can some moral or political rules be rightly imposed on all of us [...]? The answer, for proponents of public reason, is that such rules can rightly be imposed on persons when the rules can be justified by appeal to ideas or arguments that those persons, at some level of idealization, endorse or accept."
>
> (Quong 2018)

Now, unfortunately 'public' is not used consistently in the literature. Here Quong uses it to mean that justification is given "to all". However it may also be used to require that a

[6]The difference between the descriptive and the normative claim consists in the kinds of reasons that may be given in its defense. The standard view here is that in the descriptive case, reasons ought to be descriptive, whereas in the normative case, at least one premises of the justificatory argument must be normative as well.

justification is given "in public", or that it is based on "shared" reasons.[7] For now, I stick with Quong and take 'public' to mean "to all".[8]

The basic normative assumption of this inquiry is that public reason liberalism is correct in starting out with a general requirement of justification. Social order ought to be justified to all the individuals it governs. Although perhaps quite familiar, this normative claim is not some obviously accepted principle we can take for granted. As Stephen Stich states in an interview with Tammler Sommers,

> "[...] the tradition of trying to justify normative claims in a deep and foundational way, the tradition of trying to provide something like philosophical or argumentative justifications for moral judgments — this is an extremely culturally local phenomenon. It's something that exists only in Western cultures and cultures that have been influenced by Western cultures. In many cultures, and for much of human history, providing that kind of justification has played no part in normative psychology."
>
> (Sommers 2016: 289)

Consequently, a demand for public justification is itself in need of justification; a *justification of justification* so to speak. As we have just seen, this meta justification is provided by public reason liberals by reference to prior principles of freedom and equality. More precisely, they ground some specified principle of public justification in some specific principles of freedom or equality. This, however, is not exactly a normatively parsimonious starting point and creates a range of problems. Let me point out a few of them.

Firstly, to conceive of people as free and equal is not so obvious that we can take it for granted. To see this, recall the fact of social order, namely that we are all born and socialized in some web of norms, authorities and respective demands. In this light, the natural state in which human being mostly exist seems to be characterized by obligation and inequality.

Secondly, in public reason liberalism we not only need to agree on what qualifies as an appropriate principle of public justification, but also on appropriate principles of freedom and equality and how all of these things are properly related. This adds to the list of things different theorists might reasonably disagree about.

Thirdly, liberal principles may be rejected because they are not as neutral as their proponents are having us believe they are.

[7] John Rawls is well known for using 'public reason' in order to refer to reasons that are "shared" among a group of citizens. (Rawls 1997: 800) Jürgen Habermas on the other hand speaks of the "public sphere" where discourse can take place "in public". (Habermas 2008: 11-12)

[8] In *Subsection 5.1.1* I add the consideration that a justification to all is only feasible as a practice that takes place *in public*.

> "A related concern arises in light of Gaus's claim that "blameless liberty" is the moral default. This view has the virtue of directing us to simply refrain from issuing rules where matters are sufficiently vexed. But trouble emerges once we realize that many of the most controversial cases of social morality are such that the distinction between "no-rule" (blameless liberty) and a maximally permissive rule is hard to sustain. [...] The familiar moral controversies concerning the public education curriculum, same-sex marriage, gun-control, stem-cell research, pornography, among many others have at their core a dispute about whether there is a meaningful no-rule / permissive rule distinction [...]. Consequently, Gaus's proposal that blameless liberty is the default in cases of deep moral controversy will no doubt strike many of the parties to those controversies as unacceptable, an attributing to their political opposition a default victory, and thus a mere pushing around. Once again, the old worries about public reason liberalism begin to emerge."
>
> (Talisse 2014: 560-561)

Fourthly, there is the typical, but perhaps too strong requirement that all legitimate norms need to be publicly justified. In detail, the common conception of people as free and equal implies that they are *owed* a justification for being governed by any norm or authority.[9] This is a strong demand because in any actual complex and diverse society it is likely that a conclusive justification cannot be given to everyone. The main reason for this is the phenomenon of deep and reasonable pluralism, leading to persistent disagreements and controversies. We will discuss this phenomenon in *Subsection 1.2.1.* Simply put, the problem is that public justification might (to some extent) fail and where it does, public reason liberalism renders the order in question illegitimate, leaving anarchy as the only remaining option for free and equal individuals. (Enoch 2015)

This is far from a complete listing of the problems with public reason liberalism.[10] However, I do not wish to get any deeper into these discussions, because I believe we can avoid several controversial liberal commitments altogether. Specifically, I think we can do away with first principles of freedom or equality, which typically result in a requirement of public justification in respect to any kind of coercion. These are familiar ideas, but they are nevertheless controversial and can be avoided by a justification of justification that directly appeals to the justified social order as a desirable social state – or so I argue in the following.

[9] To see this, consider a typical principle of public justification: "A coercive law L is justified in a public P if and only if each member i of P has sufficient reason(s) Ri to endorse L." (Vallier 2018)

[10] For a more comprehensive treatment see Jonathan Quong (2018) and Kevin Vallier (2018).

The Justification of Justification

Why should the notion of justified social order occupy a central role in normative social theory? Is it really of any importance or is it perhaps just an intellectual game of political philosophers and theorists? What publicly justifies a public requirement of justification? In trying to give an answer, we are applying the language game of reason giving to itself by asking: What reasons justify to all that justificatory reasons should be given to each individual in favor of her social order? With the following four points I hope to be able to give a fairly straight-forward answer to what justifies the ideal of justified social order.

1) Well-Being: Justified social order is is an order in which we feel at home. This is because, if the norms we live by are affirmed by the reasons we have, we can feel free to live the lives we want, irrespective of being constrained by social order. If, on the contrary, our own reasons contradict the demands made on us by the norms we live by, these demands feel like alien dictates and their enforcement feels coercive.[11]

2) Intrinsic Value: A demand for justified social order is not limited to Western, democratic tradition. Even autocracies and theocracies may provide reasons for the norms they are built on. Thus a wide conception of justification applies wherever humans, equipped with the power to reason, stop to wonder whether they should adhere to some authority they are confronted with.

There is however a cultural difference in what is typically accepted as an appropriate justification of social order: "There is [...] an important connection between liberal argumentation and the Enlightenment conviction that everything real can in principle be explained, and everything right can in principle be justified, *to everyone*." (Waldron 1999: 229, my emphasis). In Western, democratic tradition, the individual takes center stage. In contrast to other traditions the individual is not merely to be *informed* about what justifies the norms she lives by (say that there is a God who has handed down ten commandments). Crucially, the enlightened individual must *confirm* the justification that is brought before her in light of her own reasons. "Public justification is not simply valid reasoning, but argument addressed to others [...]." (Rawls 1997: 789)

Methodologically speaking, we can refer to this way of thinking as "normative individualism". It is to say that the individual is a unique source and judge of normative claims. Thus the

[11] "A [...] fundamental interest in publicity emerges when we see that individuals' judgments often reflect modes of life to which they are accustomed and in which they feel at home. To live in a world governed by the principles one adheres to as opposed to someone else's is often, in Michael Walzer's apt simile, like living in one's own home furnished by one's own familiar things and not in someone else's or in a hotel. The interest in being at home in the world is fundamental because it is at the heart of the well-being of each person." (Christiano 2008) The reference here is to Michael Walzer (1988): *Interpretation and Social Criticism*, In: Tanner Lectures on Human Values VIII, Salt Lake City.

task of justifying social order is the task of giving reasons that can be confirmed by the well-considered reasoning of all individuals governed by said order. This is a narrower understanding of what counts as appropriate justification, focusing on the individual and her capabilities as a reasoner.

This concern for the individual lies at the heart of the tradition of social contract theory, which asks for principles of social order that everybody can agree on, as well as the democratic tradition of government *for* the people, realizing popular sovereignty. Hence I assume that having justification in a narrow, individualistic sense is of intrinsic value, at least to people who see themselves as the ultimate judges of what is good and right. Accordingly, not any justification will do. What people actually want is a justification of social order that is addressed to their own well-considered reasoning.

This is one core reason to theorize about justifying social order. Of course, as with John Rawls' cultural foundation of liberalism[12], this reason may not speak to all human kind. I accept this limitation and comfort myself and the reader with the fact that there are more reasons to be considered.

3) Stability[13] A great advantage of a publicly justified norm, in particular if its justifiedness is common knowledge, is its stability. Such norms are stable because, one, even in absence of sanctions, people have motivational resources to adhere to the norm. Two, justified norms are stable because they rule out the possibility of norms based on false beliefs. Unjustified norms may persist because people hold symmetric false beliefs about what is generally expected in society. Such norms are public bads and inherently unstable, since they might crumble like a house of cards as the wrong beliefs are exposed. Now, of course stability may also be achieved by means of sanctions. This brings us to the consideration of efficient norms.

4) Efficiency: Justified social norms require little enforcement and are thus very likely to be more efficient. As an example, think of a norm that requires you to pay taxes. If the state provides you with good reasons for paying taxes, i.e. that the money is used to fund public goods such as roads, schools and parks, this will increase your motivation for following the norm and paying your taxes. If on the other hand, you think that the state is wasting or outright embezzling tax money, this will subtract from your motivation to follow the norm. In such a scenario the state probably has to spend more and more resources on getting you to pay your taxes and these enforcement efforts may eat up a significant proportion of the money the state is trying to collect in the first place. Essentially, enforcement is costly and has a limited

[12] "The third feature of a political conception of justice is that its content is expressed in terms of certain fundamental ideas seen as implicit in the public political culture of a democratic society." (Rawls 1993: 13)

[13] My reflections on justified norms as stable and efficient norms are significantly informed by what Cristina Bicchieri calls "legitimate" norms. (Bicchieri 2006: 21; 23-24, 2017: 38)

potential.
This is of course not to say that giving people good reasons for compliance solves all problems. Even if norms are justified to all, collective action problems, e.g. incentives to freeride on the efforts of others, may persist. Therefore, even if all individuals have their own good reasons to generally follow some norm, they still might need an incentive (moderate level of enforcement) to actually do so in a given situation. But the general hypothesis here is that more justification contributes to lower costs of enforcement.
To actually explain why this might be so, i.e. why and how having good reasons motivates us to follow a norm, is not a trivial matter. Generally I assume that reasoning about a norm and considering it to be coherent with one's own interests, goals and beliefs results in some form of psychological attachment to that norm. A common thesis in this regard states that having good reasons in favor of some norm leads to the *internalization* of that norm. (G. Gaus 2011: 12.3) Internalized norms will be followed habitually and a violation of such a norm produces swift and strong emotional reactions. If widely achievable, such internalization would render any external enforcement unnecessary because then people would have their own internal motivation to follow the norm and make sure that others do the same.
Although the scenario of justified norms that are also internalized norms is perhaps the most desirable and most efficient state of affairs, it is uncertain whether it actually works. That is, it is uncertain whether reasoning about norms does lead to the internalization of norms. We will discuss the details and one potential alternative way of psychological attachment in *Chapter 4*.

In conclusion of these points, justified social order is a highly desirable ideal. Here I have listed four reasons why, but perhaps there are more. Admittedly, some of these advantages (i.e. stable order) may be achieved differently, however justified social order is unique in that it is likely to jointly achieve a wide range of desirable goals.
There are two further, more theory-internal reasons in favor of theorizing about the good society in terms of publicly justified social order. One is that it leads to a plausible normative approach in light of more foundational theories of normativity – i.e. in light of metaethics. This point, explicating the metaethical plausibility of public justification, will be discussed below in *Subsection 1.3.1.* Another reason in favor of starting out with a public justification requirement is that it allows for the construction of a normatively parsimonious theory that coheres with relevant empirical research and avoids some controversies at the level of normative construction. This point will be gradually explicated throughout *Chapters 2, 3* and *4*.
To be sure, my justification of justification is not necessarily convincing to all human beings or all human communities. In practical terms this is to say that there are individuals and

combinations thereof to whom the ideal of justified social order is not well-addressed because their fundamental interests, beliefs and worldviews contradict the requirement of giving good reasons to all. I accept this limitation and only claim that the ideal of justified social order is well-suited primarily to people influenced by Western enlightenment tradition, but not necessarily to all human kind.

In more philosophical terms, I have not answered Steven Wall's challenge to provide a justification of justification that "cannot be reasonably rejected". (Wall 2017: 385) But this is an impossible task anyway. Giving a justification means giving good reasons, i.e. reasons that relate to the reasoner(s) they are aimed at. It usually does not mean giving foundational reasons that cannot be reasonably rejected. Hence there is no contradiction in trying to justify justification in a non-foundational way, but rather a demand for good reasons at different levels. And the fact that any requirement of justification can in principle be reasonably rejected is not a defeater of any such approach. But it does motivate a concern for the possibility that the account defended can be self-testing in practice in that it allows for or even demands the possibility of citizens actually confirming or rejecting it. We will return to this point below in *Subsection 1.3.2.*

1.2 Problem: Complex and Diverse Society

At this point I hope that the basic concepts and the topic of this inquiry are fairly clear. To recap, social order is to be thought of as a set of social norms. These norms, or rather all those individuals holding the respective expectations that constitute the norms, make normative demands on people to act in a particular way. Such demands in turn may trigger the counter demand of justification, i.e. someone is asking for reasons, showing the normative demand to be coherent with the reasons she has. Consequently, a justified social order is a set of norms that has been shown to be coherent with the reasons of all governed by that order.

I further hope to have shown that the ideal of justified social order is itself justified to most communities because it is getting at a highly desirable social state, namely an order we want and that works well.

In this section we turn to the core problem of trying to explain how social order could be justified to all. This problem arises when we focus on the most challenging scenario for any theory of justifying social order: a complex and diverse society. At the end of this section we will arrive at a refined statement of our guiding research question, replacing the general concern for publicly justified social order stated at the very beginning of this chapter.

1.2.1 Reasonable Pluralism and the Diverse Society

The task of justifying common norms within some group need not be problematic. Maybe the reasons in favor of those norms are rather obvious and shared by all – here you might think of a norm that prohibits smoking on a public bus or cutting down trees in a public park – or because the group in question is very homogeneous respective to the social practice in which those norms are embedded. This could be the case if all individuals share the same culture which grounds the norms in question, e.g. the ten commandments among a group of Catholics. Hence, justifying social order may at times be a fairly straight-forward practice that does not require any refined theoretical efforts. Such scenarios inform us about how justification of social order works, but they are not the primary concern of this inquiry. What we are interested in here are complex orders coordinating social life in large, anonymous and diverse societies. How "large" exactly does not really matter. Typically what I have in mind here are entities such as cities, nation states or supranational bodies (such as the European Union) and their norms. The crucial characteristics of such bodies are, one, that there is a complex system of formal and informal norms, including anything from basic rules of everyday conduct all the way to political procedures and constitutional rights. Two, that the group in question has not created this complex order by means of an original contract, but each individual is simply born into this set of given norms. Three, the group cannot simply come together in order to confirm the existing norms or agree on new ones because of the size of the group and the diversity of their views, beliefs and goals.

Deep and Reasonable Pluralism

In reference to John Rawls' "circumstances of justice"[14] we may refer to these scenarios of diverse individuals living under a complex and given order as *the circumstances of justification*. Among the challenges associated with the circumstances of justification, dealing with reasonable and deep pluralism has received the most attention. I am not entirely sure why this is the case. One reason may be that the problem of pluralism challenges the possibility of justified social order, even on a conceptual level. This is perhaps the kind of challenge that tickles the intellectual fancy of most normative theorists. Be that as it may, let us try to get a better understanding of what the challenge posed by deep and reasonable pluralism is all about. Again, we turn to John Rawls:

> "Now the serious problem is this. A modern democratic society is characterized not simply by a pluralism of comprehensive religious, philosophical, and moral doctrines but by a pluralism of incompatible yet reasonable comprehensive doctrines. No one

[14] See Rawls' *A Theory of Justice*, § 22.

> of these doctrines is affirmed by citizens generally. Nor should one expect that in the foreseeable future one of them, or some other reasonable doctrine, will ever be affirmed by all, or nearly all, citizens."
>
> (Rawls 1993: xviii)

The conception of a "reasonable comprehensive doctrine" plays an important role in Rawls' theory. Ultimately, his aim is to show us that all reasonable doctrines can converge on an overlapping consensus on democratic institutions and a conception of justice. But since we are not interested in Rawls' theory of justice, the details of his vocabulary do not concern us here. What is important at this point is, firstly, the assumption that people hold "incompatible doctrines" – i.e. different world views and self-conceptions – that lead to disagreement on what norms they should live by. Secondly, these disagreements are *deep* and *reasonable*. They are deep because they extend to all levels: everyday norms, politics, constitutional issues and metaphysical matters such as religion and normative theorizing. The important consequence of deep pluralism is that disagreements on one level cannot be resolved by seeking agreement on a different level. Reasonable pluralism is the phenomenon that disagreement is persistent no matter how good the intentions of the discussants or how long their discussions are. Essentially, reasonable disagreement is not due to a fault in reasoning, non-ideal circumstances or defects in character.

Diversity

Of course it is not necessarily the case that a given society is characterized by deep, reasonable pluralism. Some actual society might be made up of people that can agree on most things or people that cannot agree on anything. However, the normative theorist is interested in those scenarios that are located between these two extreme cases. That is, we are interested in societies that are pluralistic but not too pluralistic. That is to say that the relevant cases are characterized, on the one hand, by deep and reasonable pluralism so that there is some challenge to be taken on by the theorist in the first place. On the other hand, there must be at least some common concerns, typically a common interest in cooperative social order. For if there were no common concerns at all, any attempt to start theorizing would be as futile as building a swimming pool in the clouds.

As discussed above, the benefits of social order render it fairly likely that a given society can at least agree on having some set of stable norms, facilitating peace and cooperation. What then makes the existence of pluralism a likely scenario? There are at least two prominent causes. One is that, in large and complex societies, the issues at hand are complex and difficult to solve. Thus, as the bounded reasoners we are, it is simply part of the nature of discussing such

complex and difficult matters that the results remain inconclusive. This is what Rawls calls "the burdens of judgment" and he provides an extensive lists of things that stand in the way of reasonable agreement on social order.[15]

The other cause of deep and reasonable pluralism is diversity. Diversity denotes the fact that most large societies are inhabited by individuals with very different perspectives on themselves and the world, typically resulting from different cultural or socio-economic backgrounds. These differences in perspectives can lead to very deep disagreements that are not about the appropriate principles, rights or values, but the world they apply to. Gerald Gaus (2016) dedicates a whole book to this phenomenon and how to deal with it in normative theory. Here is an illustrative example of his:

> "Some of our deepest and most intractable disputes are not about values or principles of justice, but about the world to which these principles apply. The most obvious instance is the longstanding and persistent struggle concerning abortion rights. Advocates of such rights see the case as decisively about fundamental rights of personal autonomy; opponents of abortion rights are depicted as having little sensitivity to a woman's claim to control her own body. But this by no means follows, and often is simply not the case; opponents of abortion can be deeply devoted to such autonomy, but not in cases where it entails overriding another's right to life. And, of course, in the abstract, most advocates of abortion rights would also draw back in such situations. The dispute is centrally about the social world to which the principles of autonomy and the right to life apply: the two social worlds do not have the same set of persons, and so even perfect agreement about abstract principles of justice would not resolve the dispute."
>
> (G. Gaus 2016: 162-163)

In summary, under the circumstances of justification – diverse individuals governed by a complex given order – the possibility of achieving justified social order is uncertain. This is because

[15] "a) The evidence – empirical and scientific – bearing on the case is conflicting and complex, and thus hard to assess and evaluate. b) Even where we agree fully about the kinds of considerations that are relevant, we may disagree about their weight, and so arrive at different judgments. c) To some extent all our concepts, and not only moral and political concepts, are vague and subject to hard cases; and this indeterminacy means that we must rely on judgment and interpretation (and on judgments about interpretations) within some range (not sharply specifiable) where reasonable persons may differ. d) To some extent (how great we cannot tell) the way we assess evidence and weigh moral and political values is shaped by our total experience, our whole course of life up to now; and our total experiences must always differ. [...] e) Often there are different kinds of normative considerations of different force on both sides of an issue and it is difficult to make an overall assessment. f) Finally, [...] any system of social institutions is limited in the values it can admit so that some selection must be made from the full range of moral and political values that might be realized." (Rawls 1993: 56-57)

under these circumstances, we expect persistent reasonable disagreement about the right norms to live by. How can any norm ever be justified to all under these circumstances?

1.2.2 The Problem of Justification in Theory

The challenge to the idea of justified social order posed by the phenomenon of deep and reasonable pluralism seems daunting. But, given that the immense benefits that result from living under a stable justified social order stated above do in fact materialize, we do not need to give up on this ideal. In fact, the prospect of living under conditions of stability, peace and prosperity should provide quite weighty reasons to most people in favor of accepting the reign of a set of social norms.

But this in itself is not a very interesting claim. With the "fact of social order" I have already stipulated that there is an obvious public justification for social order per se. So normative theory should have more to offer. It should allow us to make some headway on the question that really matters to actual people in a given society: *What* social order is (most) justified to us?

Unfortunately, the normative theorist is not in a good position to answer this more interesting question. To see this, first consider the task that giving an answer would amount to: In order to show that some norm N or set of such norms is justified, one would have to show that all individuals governed by that norm have conclusive reasons in favor of N. Having "conclusive" reasons in favor of N means that one has reasons that favor N, while these reasons are undefeated by other reasons speaking against N. More simply put, the task for the theorist is to show that all individuals governed by N strictly prefer N to not-N.

Second, consider that there are at least three obstacles standing in the way of the normative theorist who is trying to fulfill this task.

1) Knowledge

The task of the normative theorist requires knowledge about individual preferences, but she does not have access to these preferences. As a theorist she is not an expert on what people actually believe, think or want. In small groups it might be plausible to assume that this information is easily attained, but in large-scale societies this is very difficult. Maybe politicians, journalists and pollsters have some idea of what the people actually want. The normative theorist on the other hand seems to be quite ill-prepared for joining their discussion. Neither her methods nor her professional experience qualify her to make well-founded claims on people's preferences. This is the knowledge problem of justifying social order.

For the sake of argument let us assume the knowledge problem could be overcome by means of a study surveying people's preferences and reporting them back to the normative theorist.

Including surveys in normative theorizing is not unheard of.[16] This, however, leads us to the problem of idealization.

2) Idealization

Let us go back to the idea of a social contract. The point of having a contract is to end up with a situation where some social arrangement – say government – is justified and can legitimately make use of its powers because it is backed by the agreement of the governed. In order for this idea of justification by consent to work, the choice has to be *well-considered*. That is to say, the contracting parties have to be well-informed, act voluntarily and they should have given their choice a sufficient amount of thought. Otherwise, your consent does not have any normative significance, which is most obvious if you consider a forced choice where somebody puts a gun to your head. Therefore, agreement does not provide any grounds for justification if these conditions are not met. So for the justificatory argument to work, we need an agreement under ideal conditions of voluntary, well-informed and well-reflected choice. Obtaining these ideal conditions is the idealization problem of justifying social order.

At this point you may wonder whether there are conditions of well-considered choice which, one, warrant the presumption of a normatively significant choice and, two, can be fulfilled in principle by ordinary citizens. I think there are. They can be found by following Gerald Gaus' pointer to a context-dependent standard of sufficient deliberation. His argument starts out with rejecting the image of a perfectly rational reasoner:

> "full rationality is an extravagant assumption which only disappoints, for it succumbs to the very indeterminacy that it seeks to avoid. A reasonably reflective real rationality is the most we can demand of others, and what is reasonable to demand of others depends on the nature of the social practice in which the appeal to reasons occurs."
>
> (G. Gaus 2011: 258)

Gaus is right. Our reasoning will never be perfect. And we can allow for these imperfections because for our purposes it is sufficient that people can do things for good reason. That is, individual reasoning need not be flawless and in many everyday situations people might as well do things based on bad or no reasoning at all. What is crucial is that they also have the capability to sit down, discuss some issue and then make a sufficiently voluntary, informed and reflected choice, whereby the respective sufficiency standards depend on the choice at hand and its social context.

[16] Consider for example David Miller (2003), who we will briefly discuss in *Subsection 3.3.1.*

It does for example make a huge difference if you are about to buy some tomatoes at your local supermarket or if you are about to take your vows at the altar. If you are applying the sufficiency standards of the former situation to the latter, you might end up getting married drunk in Vegas to someone you just met. Which does of course not imply that you have made the wrong choice or that you will be unhappy. It simply means that you have made your choice to get married based on insufficient information and reflection, relative to the appropriate standards for such a situation. It is thus not a well-considered choice based on *good* reasons.

Just as in the case of getting married, the sufficiency standards for making well-informed and well-reflected choices in the case of choosing norms of social order are quite high. In both cases you should be very well-informed about the options that are available and have carefully considered the consequences because it is likely that the choice you are making holds for a long period of time and will have severe impacts on your quality of life.

Whether a choice is sufficiently voluntary also has a contextual component. A choice is clearly involuntary if the outcome is directly a function of what other people want you to choose. Say if a person puts a gun to your head and says *Do X!* and therefore you do X. Nevertheless, it is a fact of life, even more so of social life, that we cannot do whatever we want and it might be the case that what others want indirectly constrains our freedom of choice without contradicting sufficiently voluntary choice. For example, if you were to live in a community made up of mostly Christians, you might only be able to choose among norms that are consistent with the Christian faith. But if these are simply the circumstances of life you happen to find yourself in, I do not see why your choice within the Christian society could not still be sufficiently voluntary. After all, the limitation of your choice set is rather a contingent circumstance and not forced on you by someone who wants you to behave in a particular way. Hence, just as in the case of sufficiently well-informed and considered choice, sufficiently voluntary choice is context-dependent.

Let us now move on by again assuming for the sake of argument that the problem at hand – the problem of idealization – can be overcome. To this end, imagine that through some awesome piece of technology, the normative theorist were able to extract the well-considered preferences on social order from the minds of all individuals and combine them in one data set. Then, you might presume, the normative theorist would be in a position to lock herself in with this data set for a few weeks and finally return to the public with a grand theory of justified social order for all. However, this cheerful scenario is also unlikely to materialize due to the problem of inconclusiveness.

3) Inconclusiveness

What would be the characteristics of the complete set of well-considered individual preferences? In principle there are two extreme cases here. One is the worst-case scenario that no conceivable norm or set thereof will cohere with individual preferences. That is, there is no conceivable norm that can be justified to all in light of the reasons they have. The other extreme scenario would be a situation where all preferences would converge on one most preferred norm or set of norms.

The fact of social order implies that the first extreme case is very unlikely because the immense benefits of social order provide good reasons for most combinations of individuals to accept some set of common norms. The second extreme case is very unlikely under the circumstances of justification and the assumption of deep, reasonable pluralism. This is because, per definition, reasonable and deep pluralism means that existing diversity is also reflected in people's well-considered preferences. So we should expect well-considered preferences to also reflect diversity.

Hence, the most likely case is that there are, one, no conceivable norms that are most preferred by all, two, some norms that are justifiable to some but unjustifiable to others and, three, some norms that are more or less preferred by all. That is, it is likely that there is a set of norms that might be more or less justifiable to all individuals, but there is no obvious way of choosing between these norms. This is the inconclusiveness problem of justifying social order.

Gerald Gaus (2011: 323-325) has captured this problem by speaking of a "socially eligible set" of justifiable norms. In respect to this set (and apart from some abstract principles) he thinks that theories of public justification remain inconclusive. In defence of this point he claims that in order to select from the eligible set of justifiable norms we would need a uniquely justified procedure for doing so. However, such a procedure does not seem to exist. (G. Gaus 2011: 19.1) Of course we could simply impose a utilitarian principle of maximizing overall preference satisfaction or another method of preference aggregation. But all such steps would be highly controversial. Not least because Kenneth Arrow (1951) has famously shown that there is no method of preference aggregation that does not violate certain intuitively reasonable criteria.

Gerald Gaus' theories will continue to accompany us in the following chapters. For now let us conclude that justified social order is a demanding and perhaps even too demanding ideal for a society characterized by the circumstance of justification and deep, reasonable pluralism. Furthermore, in normative theory we are facing the peculiar problem that we do not have access to a core building block of normative theorizing: well-considered individual preferences of social order. And even if we did, it is likely that any set of such preferences would not be conclusive.

Guiding Questions

In light of these challenges, we can specify the guiding questions of this inquiry. I started out this chapter with a broad question of public reason liberalism: How can the social rules that structure and coordinate the actions within a given community be justified to each individual? Now we can ask more precisely:

(1) Under the circumstances of justification, how can social order be justified to each individual governed by its social norms?

This is the *possibility* question. It asks for a plausible conception of justified social order under the circumstance of justification and the associated challenges outlined above. Besides the possibility question, there is the *content* question:

(2) What is the *content* of justified social order for some given community?

This second question is asking what substantial normative claims can be made from the standpoint of normative theorizing and thus provides much of the motivation for engaging in these kinds of inquiries in the first place. The crucial thing about the possibility and the content question is to keep them apart because even if the first one can be answered, this does not mean that the second one can be answered as well.

1.3 Approach: A Guiding Justification Principle

How can we hope to be able to answer the two questions just posed? Well, in order to answer a question one needs to know, firstly, what exactly the question is asking and secondly, what it would mean to answer it. In respect to the former, I have already spilled some ink on clarifying the core concepts of the two questions in the first section. In respect to the latter, in this section I follow a custom of public reason theory and provide a justification principle. This principle further specifies what it means to answer question (1) and thus guides the following inquiry.

Things are more complicated in the case of question (2). Due to the problems discussed in this section and problems that will become apparent in *Chapter 2* and *Chapter 3*, I think we cannot properly address question (2) in normative theory. Thus in *Subsection 3.2.3* I suggest a procedural restatement of question (2) that will receive at least a partial answer in *Chapter 5*.

Now, before considering our guiding justification principle, we need to reflect a fundamental assumption of this inquiry that we have already touched upon above: normative individualism.

1.3.1 Normative Individualism

As we have seen above in *Subsection 1.1.2*, justification in the narrow sense rests on individual affirmation. The methodological approach reflecting this individualistic perspective is "normative individualism". Its core claim is that the evaluation of social states can ultimately be grounded in individual mental states (desires, wants, evaluations, etc). This claim, however, may be challenged. Someone may object that ultimately it does not matter what people think or believe. What allegedly matters is what is right and true. Such objections lead us into the field of metaethics where questions about the nature of normativity and the validity of normative claims are discussed.

Perhaps the most widely discussed question in metaethics concerns the ontological status of evaluative claims. More simply put, the question concerns what it means to say things such as: "Skinning babies *is* wrong!" An "objectivist" answer to this question roughly holds that such value statements are right or wrong in virtue of some fact that is independent of what people happen to feel, think or believe. A "subjectivist" answer to this question holds that, if such value statements are right or wrong in virtue of some fact at all, this is a fact about individual mental states (what people happen to feel, think or believe).

These different positions about the ontology of normativity have led to a complex and well-rehearsed debate in which objectivists insist that their theories can account better for important normative intuitions and the view of human beings as autonomous agents.[17] If, for example, you could show that skinning babies is wrong because there is an objective fact – i.e. some intrinsic feature of the action in question – corresponding to it, you could claim that it definitely *is* wrong to skin babies, independent of what is going on in people's minds. And this is fairly close to how we think about the matter in everyday life. We usually think that skinning babies is just categorically wrong, no matter what. The typical subjectivist stance is that the most intuitive or simple explanation is not necessarily the best one and that in the end objectivists are stuck with a problematic ontology, assuming the existence of "queer" or "mysterious" entities.[18]

Normative individualism is geared toward subjectivism because normative individualism grounds normative claims in individual mental states. As you might have guessed from my line of argumentation so far, I consider myself a subjectivist and I presume that most contemporary theorists associated with the tradition of social contract theory are also subjectivists. Nevertheless, as John Rawls (1993: 95) has argued, theorizing about justified social order based

[17] See for example Joseph Raz (1999) or Christoph Halbig (2007).

[18] For the probably best-known classic argument against objective normative ontology see John Mackie (1977: Chap. 1). For a more recent charge see Peter Stemmer (2008, 2017).

on normative individualism does not and need not assume the objectivist to be wrong.[19] It is rather an effective way to move forward in face of persistent metaethical disagreements.

To see this, note first that normative individualism does not claim that the objectivist is wrong. She may continue to search for a convincing account of an objective ontology of fundamental values *and* a way to establish what they demand of us.[20] If she does so successfully, her account will automatically be reflected in a regime of individually justified social order because then her account will be well-justified to most. But as long as the debate in metaethics rages on, we need a way to move on at the level of normative social theory.

Note, second, that an account of justifying social order based on normative individualism is a practical and plausible way to do this. It is "practical" because it is about giving reasons to everybody – including the objectivist – for accepting some social arrangement. Given the fact of social order, we do need to settle on some social arrangements. Short of violence, deception or manipulation, seeking solutions that are justified to everyone seems to be *the* standard practice of moving on in the face of persistent disagreement.

It is "plausible" because it is difficult to deny that human beings are a source of practical reasoning and evaluative thinking. More precisely, it is difficult to deny that human beings have connotative mental states: they want or desire certain things and certain things more than others. These desires and wants can be and often are the object of practical reasoning. That human beings have such capacities of practical deliberation about what they should do and related mental states of wanting or desiring something is obvious. Further, there is a fairly straight-forward evolutionary explanation as to why human beings have these capacities. Namely that there is a tremendous advantage in being able to solve practical problems of individual and group decision-making while taking into account future outcomes. In light of these facts, it is difficult to deny that the human mind and its connotative states is a unique source of normativity.[21] To be sure, it does not prove that there might not be other sources. But given that the individual mind is the only source we can all agree on that is there, it seems highly plausible to ground social normative claims in individual normative claims. How else

[19] Also note that the classical defenders of social contract theory Thomas Hobbes, John Locke and Jean-Jacques Rousseau, entertaining conceptions such as natural rights and a general will, had strong objectivist connotations.

[20] "An understanding of "morality" entertained by many people is that it refers to a realm of normative facts that in some sense concerns how things "really are" and which are independent of agents' "subjective perspectives." Moral judgments are beliefs about these "objective" facts. On this notion of morality, that "Alf's action was wrong" is a fact about Alf's action that is independent of what he can see as a reason to hold that it is wrong. Nothing I say in this work is inconsistent with such a view of morality and truth. Those who hold it can continue to make such judgments of others without rejecting the analysis." (G. Gaus 2011: 229)

[21] What I am pointing to here is what one might call "an anthropological perspective to metaethics". That is metaethics informed by scientific insights into evolved human capacities of practical reasoning, evaluating and desiring. For a philosophical explication of this approach see Peter Stemmer (2016) and especially Stefan Fischer (2018: §10), who has significantly informed my brief metaethical reflections.

could we move on?

In conclusion, this brief excursion into the field of metaethics was meant to show that a requirement of public justification, based on normative individualism, is plausible because it is based on things that clearly exist: individual reasoners and their connotative states. And although I identify myself as a subjectivist, none of the above claims that objectivism is wrong. In a justified social order and under the reign of public reason, the objectivist can remain an objectivist and social order has to be justified in light of her reasons as well. Further, under such a framework, her objectivist reasons will become more and more influential if she successfully convinces more people of her position. Essentially my approach based on normative individualism is geared more toward subjectivism rather than being neutral, but at the same time it gives objectivists a fair chance to prove their point and it allows all of us to move on, irrespective of persistent disagreement on all levels.

1.3.2 The Justification Principle for Social Norms

We have at last arrived at the formulation of the principle of justified social order that will guide the following inquiry. I call it *the justification principle for social norms* (JPN) and it reads as follows:

> **JPN**: A social norm N is justified to an individual i in society S governed by that norm to the extent that N being a positive norm in S is coherent with i's preferences, given that
>
> 1) i has formed well-considered preferences on social order,
> 2) N being a positive norm in S is strictly preferred by i to having no social norm governing the domain of N in S,
> 3) i is at liberty to openly reject the JPN in S.

The basic idea behind the principle should be obvious enough. The task of justification is to show to an individual that the norm in question is in accordance with her well-considered preferences. If that person were to ask, *Why should I follow a norm of treating everybody with equal respect?*, the justification of that norm would consist in providing an argument, thus linking the norm and that person's preferences. By implication, justifying a norm to some group of individuals would involve the same task relative to every individual member. This captures our simple model of two sets from *Subsection 1.1.1* and the ideal of justified social order argued for in *Subsection 1.1.2*.

Some Technicalities

Instead of talking about "preferences", others have used the notion of having "sufficient" or "conclusive" reasons. (G. Gaus 2011: 13.3) I believe essentially what is meant in both cases is identical. But talking about preferences is the more precise way of putting it. Having a preference for social state X means that I prefer X to social state Y all things considered. I still might have good reason in favor of Y, but preferring X indicates that I think that I have better reasons for favoring X. The notion of preferences thus highlights that people having good reasons for some norm is not decisive. Decisive is what they prefer after reflecting upon the different reasons they have. Preferences are considered expressions of one's balance of reasons respective to some alternatives to choose from.

The formulation of "N being a *positive norm* in S" is meant to stress that what matters is the fact or the imagination of N actually governing i's social life. The bare content of N alone is not a sufficient basis for the justificatory argument because it would ignore the actual social circumstance and the consequences resulting from the implementation of N.

A somewhat unorthodox feature of the JPN is that the coherence relation between social norms and individual preferences proposes a gradual rather than a definitive standard of justification. More precisely, according to the JPN justification is a function of *the degree of coherence* between a social norm N and individual preferences. Of course a simpler, dualistic model would try to draw a dividing line between the justified and the unjustified. But although I also enjoy the parsimony of simple models, I think in the case of justification understood as reason giving, any attempt of coming up with a simple on/off model would be hopelessly unrealistic. This is because, especially in light of the complexities and problems discussed in the previous section, the reasons we have in favor and against social norms are also complex.

As an example, think of someone preferring norm X over norm Y and norm Y over norm Z. Relative to that person, X would be perfectly justified, for X is her most preferred norm. In respect to Y things are already more complicated. Y is somewhat coherent with her preferences, for she prefers it to Z. But there is also a contradiction between Y being the a positive norm in S and her preference of X over Y. And what about Z? It is her least preferred norm but that is not to say that there are not weighty reasons favoring Z as well. For example, i may not like Z, but still prefer it to not having any norm in place at all. Now add thousands or millions other individuals and their preferences.

The point is that favoring, disfavoring and the justifications thus grounded come in different degrees and form a complex relational web if we consider more than one individual. Placing an all or nothing point into this range of justification would be an arbitrary manipulation. Thus I believe we need a gradual standard of justification.

Well-Considered Choice

Condition (1) of the JPN states that preferences ought to be *well*-considered preferences. This requirement reflects that, as elaborated in *Subsection 1.2.2*, according to a narrow understanding of justification, justificatory reasoning has to meet certain standards of reasoning well. That is, reasoning should be sufficiently voluntary, well-informed and well-reflected. This is difficult to pin down more precisely. Roughly speaking this is at least to say that individual reasoning should not be distorted by extortion, deception, missing information or flawed reasoning.

Of course one would expect that a theory of justified social order will eventually have more to say on the precise sufficiency standards for the justification of social norms. Such expectations are understandable, yet I will resist giving a precise explication in theory for two reasons. One, having already agreed with Gerald Gaus that such sufficiency standards are context-dependent, any attempt to give a general account seems questionable. Two, I worry that any attempt to formulate a precise conception of sufficiently well-considered preferences will end up like the attempt to give a more precise account of 'knowledge' understood as true and justified belief: It will end up in an endless back and forth of proposals and counterexamples.[22]

Alternatively, in *Chapter 5* I subscribe to a procedural answer, which roughly states that citizens can be reasonably expected to satisfy *Condition (1)* if their preferences have been shaped by high-quality public deliberation in the public sphere.

The State of Nature Condition

The range of justification has a lower limit: Any justified norm should at least be preferred to not having a norm in place at all in the domain of that norm. This is a common-sense restriction because if someone disfavors not-N over N all things considered, N is not justified to that person.[23] Therefore, *Condition (2)* of the JPN establishes that the minimum threshold for a norm to be justified to any degree is its superiority to the state of nature.

Admittedly, there are good reasons to think that *Condition (2)* is both too wide and too narrow. It is presumably too wide because it allows social norms that only hold due to strong power asymmetries to achieve some degree of justification. That is, N could allow for someone being a wage slave, while that person might prefer N to having no regulation at all because she is surrounded by very powerful individuals and is afraid of ending up even worse, for instance as a slave who is not paid at all. Of course, one can argue that in such a scenario the wage slave also has good reasons to agree to this arrangement and thus N is in fact somewhat justified

[22] Here I am of course referring to the debate in epistemology on whether 'knowledge' is correctly defined as justified true belief, initiated by Edmund Gettier (1963).

[23] Gerald Gaus has proposed a similar baseline requirement but in his case the state of nature is a "blameless liberty" to do whatever you want as long as you do not the violate basic rights of others. (G. Gaus 2011: 316-317) On my account, the state of nature baseline is simply to not have the rule in question in place.

to her. A standard objection to this kind of account would be that it equates justification with rational choice while ignoring that we want justification to mean more: A truly voluntary choice, anchored in a person's own goals, values and vision of the good life. I am drawn to both lines of argumentation and thus incorporate both into my normative model of justified social order in *Section 4.2*.

You might, on the other hand, think that *Condition (2)* is too narrow because it could allow individuals to veto a wide range of norms until there is not much left of what otherwise might be considered a formidable social order overall. This problem arises when we think of social order as a whole package of norms and then imagine individuals who pick out norms that do not meet *Condition (2)* according to their preferences. Pointing to the JPN, many individuals may then veto social norms because they prefer having no norm in place in this domain. For example, a wealthy person may prefer no regulation to any form of mandatory public schooling and financing thereof. Many such vetoes may come in until there is not much left of social order.

The response to this worry is of course that in fact social order is mostly a package deal. Individuals usually consider whole sets of norms that constitute some social practice. The actual choice then is between accepting all norms and participating, or non-participation. For instance, if you choose to enter a supermarket to buy some groceries, you cannot cherry pick which norms involved in this practice you adhere to and which you reject. That is, you cannot decide to adhere to the *First come first serve!* and the *Present all items at the cash register!* norm but reject the *Pay in the accepted currency!* norm. Effectively, social order as a whole would break down if *Condition (2)* were to be understood as authorization to reject singular norms wherever some individual might prefer having no norm in place. Thus in most actual cases, a person thinking about some norm N and *Condition (2)* needs to take into account that N is likely to be an inseparable part of some social practice. This means that a person who has conclusive reasons in favor of some practice will have pragmatic reasons to accept some norm N that is part of that practice, even though that person would not prefer N to not-N on its own. This kind of pragmatic package deal justification is perfectly fine in most cases where we do not care much about the norm in question. It might however break down if the norm in question is defeated by important personal reasons. We will return to this point in *Section 4.2.2*.

Self-Testing

Condition (3) is necessary because the JPN is the kind of "reflexive" standard that applies to itself. (D'Agostino 2013: 132) "Some standards apply to themselves, and then they either meet the standard or they do not." (Estlund 2008: 54) The JPN falls into this category of standards

because it builds on, one, normative individualism, i.e. the basic claim that social states can ultimately be evaluated respective of individual mental states, two, on an ideal social state of justified social order and, three, on the actual content of the JPN, specifying what the ideal requires. Hence, the ideal of justified social order must, according to normative individualism, also be publicly justified and it would be odd if the standards for the meta justification were different from the standards specified by the JPN. Then we would have to explain why the meta justification can be a good justification while ignoring the very standards of good justification specified for the case of regular justification.
In short, the substantial components of our account of public justification themselves require public justification. I have provided such a meta justification by arguing for justified social order as a highly desirable ideal in *Subsection 1.1.2* and by arguing that it is metaethically well-founded in *Subsection 1.3.1.* Nonetheless, I have also acknowledged that these justifications of justification may be reasonably rejected by some. According to normative individualism, such rejections are indeed a fundamental problem of our account of public justification if it were applied to actual societies. Essentially an account of public reason that forces a non-publicly justified ideal on society is authoritarian by its own standards.
Therefore, we cannot simply impose the reign of our account of justified social order from normative social theory onto society at large. Rather, *Condition (3)* requires citizens to be in a position where they can form and voice well-considered preferences, not only on social norms, but also on requirements or ideals of justified social order. Consequently, any theory of justified social order hoping to satisfy the JPN should explicitly incorporate the empowerment of citizens to reflect upon and even rejected the JPN. This way we ensure that normative theorizing is not authoritarian, but self-testing.[24]

The matter of self-testing, as many other points we have come across in this first chapter, will continue to occupy us in what is still to come. The JPN in particular will guide us time and again as we get to the positive part of this inquiry and start developing our own theory of justified social order.

1.4 Concluding Remarks *Chapter 1*

This also concludes the first chapter. As in all of the following chapters, at this point I offer an argumentative summary of what has been said.

1) We start out with a standard question of public reason liberalism: How can the social rules that structure and coordinate the actions within a given community be justified

[24]Seyla Benhabib (1990: 340) proposes a similar solution to the problem of justifying the justificatory practice of ideal discourse by suggesting that opponents of ideal discourse may voice their discontent *within* ideal discourse.

to each individual? The goal of this chapter is to understand what this question is all about.

2) The fact of social order: Human beings generally do *and* want to live in stable communities, which requires a common set of rules, coordinating expectations and regulating interactions.

3) A simple model of two sets: Social order is a set of social norms that governs the behavior of a set of individuals. "Social norms" are rules that structure the *interactions* of a set of individuals and make general demands to do or not to do something.

4) These normative demands implied by a social norm may trigger a demand of justification. Justification in respect to social order is the practice of giving reasons to someone governed by a norm in favor of the normative demand implied by that norm.

5) In order to get to a normative perspective on social order we adopt a principle of public justification, requiring norms to be justified to all they govern.

6) The public justification requirement calls itself for a public justification, which consists in showing that justified social order is a highly desirable social state. It is highly desirable because it promises an order that is pleasant, valued, stable and efficient.

7) The central challenge to the requirement of public justification emerges under what I refer to as *the circumstances of justification*. This denotes the most relevant scenario for our inquiry, namely societies made up of diverse individuals governed by a complex given order.

8) Most prominently, diversity takes the shape of *deep and reasonable pluralism*. That is, persistent disagreement on all levels that is not due to some default in (collective) reasoning, lack of information, time or sincerity.

9) In light of the circumstances of justification, the guiding questions for this inquiry are (1) *How can we conceive of publicly justified social order as a possible ideal?* and (2) *How could some actual society pursue this ideal of justified social order?*

10) My understanding of public justification rests on normative individualism; the claim that the evaluation of social states can be grounded in individuals' mental states. Normative individualism is *plausible* because it is based on things that clearly exist: individual reasoners and their connotative states. And it is *conciliatory* because it does not reject objectivist positions.

11) Finally I state the guiding *justification principle for social norms* (JPN): A social norm N is justified to an individual i in society S governed by that norm to the extent that N being a positive norm in S is coherent with i's preferences, given that (1) i has formed well-considered preferences on social order. (2) N being a positive norm in S is strictly preferred by i to having no social norm governing the domain of N in S. (3) i is at liberty to openly reject the JPN in S.

As a final note in this chapter, let me recap some basic choices in terms of how to frame the inquiry. The question asking about the possibility and content of justified social order is a standard question of public reason and social contract theory. Nevertheless, I have departed from public reason liberalism as it is commonly perceived. First and foremost, I have not made any normative assumptions, such as basic principles of freedom or equality, other than the principle of public justification itself. This is because I tend to see further normative assumption as unnecessary baggage, standing in the way of attaining a sober understanding of the facts on the ground – as far as this is even possible from the abstract perspective of normative theory. Consequently, I do not see justified social order as a necessary requirement or something people are owed, but as an ideal they might aspire to. A further peculiarity resulting from the JPN is that justified social order comes in degrees. For me it seems obvious that this must be so in light of the complex relations between individuals and their social orders. The gradual perspective also goes well with stipulating justified social order as an ideal because ideals can be gradually realized by individuals willing to pursue them. Last but not least I have a tendency to think that the answer to substantial problems consists in deferring them to processes or procedures. This is motivated by the problems of solving substantial questions in abstract theorizing pointed to above in *Subsection 1.2.2*.

I mention these things here because we are about to critically engage with and ultimately reject a prominent approach to the JPN (or some adjacent principle) from the realm of public reason liberalism. Engaging in such discussions, I am often haunted by the sensation that the discrepancies at hand are at least as much about the philosophical predispositions as they are about tangible arguments. Fortunately, there are also tangible arguments. Let's get to them.

Chapter 2

Against Hypothetical Choice Modeling

In the first chapter I have introduced the phenomenon of deep and reasonable pluralism that characterizes divers societies. Further, we have seen that as normative theorists, thinking about how to justify social order, we are confronted with the problem of having no direct access to people's preferences – especially not to their well-considered preferences. And even if we had this data available, there is no guarantee that the set of well-considered preferences would add up to any particular outcome. So how to proceed in light of these challenges?

Arguably the most prominent approach in the last several decades has been that of constructing models of hypothetical choice, which derive claims of good social order from arguments about what people *would* choose under the right kind of conditions, theoretically specified. Now, the most prominent and influential author making use of this approach is of course John Rawls. We will discuss his work in some detail below, whereby one challenge in discussing Rawls consists in taking into account the vast amount of literature that has already done so. I somewhat mitigate this challenge here by focusing on the methodological rather than the substantial side of hypothetical choice modeling. Consequently, in this chapter I do not engage with the entire theories of the authors I discuss. Rather, I focus on a uniting methodological feature, one might call *hypothetical constructivism*, that is typically found in contemporary theories of the social contract. I believe proceeding in this way is justified because methodological discussion tackles fundamental problems that are usually reflected in all substantial outcomes of a theory. Also, methodological issues often receive less attention and thus I hope to be able to offer criticism that has not been considered in a systematic fashion.

I gear up the critical reflections in this chapter by firstly laying out a more precise conception of hypothetical choice modeling and introducing two prominent applications of this approach in the literature. In a second step, I try to clarify the methodological nature of hypothetical choice modeling by contrasting it with the neighboring methods of thought experiments and as-if modeling, and by delving into the nature of constructivism. In a third step I present two

strands of criticism, which respectively attend to two different readings of what hypothetical choice modeling is all about. I eventually conclude that the deep methodological problems displayed in this chapter provide us with good reasons to start looking for an alternative approach - an alternative which I develop in the subsequent chapters.

2.1 What is Hypothetical Choice Modeling?

In social contract theory, hypothetical choice modeling - in the following referred to as "HCM" - has emerged as a reaction to David Hume's effective critique of classic social contract theory. Particularly John Locke presented us the social contract as a historical event, an original agreement of free man, constituting political society and eventually legitimate government.[1] In his 1748 Essay *Of the Original Contract* David Hume clarified that, while being a desirable scenario, in fact government does not rest on the consent of the governed, nor is this a realistic demand for the future. In reaction to Hume's critique, theorists have turned to the idea of hypothetical instead of actual agreement. Simply put, the idea is to accept that people have not and will not actually come together and explicitly agree on anything and then turn to the question of what people *would* agree on if they *were* in some optimal setting for finding comprehensive agreement. Sometimes this hypothetical turn is already associated with Kant's take on the social contract or his categorical imperative.[2] The most prominent and paradigmatic statement of social contract theory in its hypothetical interpretation is without doubt John Rawls' (1971) *A Theory of Justice*. Since this work has been extremely influential in contemporary political philosophy and beyond, it is not surprising that HCM as one of its core components strongly resonates in many of the following works until today. Besides Rawls I discuss in detail Gerald Gaus' theory as an instance of HCM. Further, John Harsanyi's and David Gauthier's well-known contributions are briefly considered later in the chapter. This is explicitly not meant as an evaluation of any of these substantial theories as a whole, but should provide enough evidence for the claim that HCM is indeed an important and prevailing phenomenon in normative social theory. But before we turn to any specific theory, let me outline the three basic components and features of HCM.

2.1.1 HCM in a Nutshell

Models of hypothetical choice derive claims of good social order from arguments about what people would choose under the right kind of conditions, theoretically specified. Fred D'Agostino,

[1] See John Locke's *Second Treatises of Government*, especially Chapter VIII.

[2] See Kant's remarks on the original contract (Kant, MS, AA VI: 315-316) or his statements on the categorical imperative as a thought experiment about universally reasonable legislation (Kant: GMS, AA IV: 434, 438; KpV, V: 30).

Gerald Gaus and John Thrasher summarize this idea as follows:

> "Social contract theories are a model of justification that have several general parameters that are set differently in different theories. What distinguishes contractarian theories is how they specify these general parameters. The goal of the model is to represent our reasons for endorsing and complying with some set of social rules, principles or institutions. This is done by showing that some model representatives choosers [sic] who would agree to these rules in some specified choice situation. [...] The social contract, then, is a model of rational justification translating the problem of justification (what reasons individuals have) into a problem of deliberation (what rules they will agree to)."
>
> (D'Agostino, G. Gaus, and Thrasher 2017)

So we can say that models of hypothetical choice are constructed by means of three basic elements. One, the choosing *agents*. These are abstracted and idealized representatives, representing us as actual citizens in the theoretical model. Two, the *choice situation* in which the agents are confronted with a specific choice problem under circumstances that usually differ drastically from choice problems and circumstances in everyday life. Three, there is the choice outcome in the form of *normative claims* - i.e. principles, rules or institutions - which are unanimously chosen by all agents. The goal of the whole exercise is to show that the agents would all agree on some normative claims in the choice situation as specified by the theorist. In other words, the first important goal of HCM is to provide a conclusive model. A second core goal of HCM is to show that the model reveals good reasons we have as actual citizens for also endorsing its normative claims.

Besides the three basic elements - agents, choice situation and normative claims - there are three further key characteristics of hypothetical choice modeling. The first characteristic is - not surprisingly - *hypothetical choice*. Typically, theorists idealize the capacity, available resources and motivation agents have for making rational, well-reasoned choices. They further abstract from the obstructions to ideal choice making we all face in actual social life. In short, agents are usually conceived of as ideal deliberators under ideal deliberative conditions, only concerned with things that are relevant for the specific problem presented to them by the theorist. As a consequence of abstraction and idealization, choosers and their choice are counterfactual – they are not realized in social reality, but only within the philosophical reflection in question. The second key characteristic of HCM is *deductive modeling*. This means that – similar to choice problems in game theory – once agents and their choice situation are well-defined, the choice outcome is meant to follow as an evident conclusion, without any need for empirical testing.

A third key feature of HCM are its *substantial normative claims*. This is meant to refer to the fact that theorists employing HCM do not content themselves with presenting an account of a normatively adequate choice situation for choosing justified principles, rules or institution. They further go on to speculate which principles, rules or institutions citizens would choose in that situation. And the outcome of these speculations are claims about which principles, rules or institutions can count as justified to us as actual citizens within a particular society. In other words, theories employing HCM typically seek an answer to what I have called the possibility and the content question of publicly justified social order in *Subsection 1.2.2*. The answer to the possibility question in HCM is the theoretical fiction of a hypothetical choice model. The answer to the content question consists in the normative principles derived from this fiction. The latter are "substantial" normative claims in that they make claims about the content or structure actual social orders ought to embody.
After these brief remarks on the nature of HCM we now turn to its most paradigmatic instantiation.

2.1.2 Rawls' Original Position and Constructivism

John Rawls *original position* is without doubt the most prominent instance of hypothetical choice modeling. Rawls first introduces the notion of an original position in his 1963 paper *Constitutional Liberty and the Concept of Justice*. Here, he already uses it to denote a specific, counterfactual choice situation from which – so he argues – individuals would choose his two famous principles of justice.[3] Throughout his following works – most notably in *A Theory of Justice* (1971), *Political Liberalism* (1996) and *Justice as Fairness: A Restatement* (2001) – the original position remains a core component of Rawls' theory.

The Original Position

It should firstly be noted that Rawls' original position is meant to solve a more specific and arguably more narrow problem than the general problem of justified social order as presented in *Chapter 1*. Rawls is not concerned – as I am – with the justification of *any* social norm relative to *any* set of individuals governed by it. Rather, he restricts his justificatory problem to the question of constitutional justification within a society made up of reasonable and pluralistic democrats. Further, "constitutional justification" here does not mean that Rawls is concerned with the justification of a set of constitutional laws as it might exist in political reality. The question he is concerned with is the normative meta problem of justice with respect to such a

[3]Interestingly though, in this early paper the notion of an original position is not yet accompanied by the idea of a veil of ignorance. Rather, here the original position is much closer to classic social contract theory, where the normative virtues of the contract are simply ensured by its binding nature and a unanimity requirement. Only in a later paper on *Distributive Justice* (Rawls 1967) does he introduce the idea of a veil of ignorance.

set of constitutional laws. From this perspective, the most basic question of political philosophy is not, *What are the correct constitutional norms?*, but rather, *What are the correct standards that should govern discussion and choice of constitutional norms?*.

Rawls' construction of the original position as a solution to this problem starts from one core normative principle: *fairness*. His procedural solution to the problem of choosing principles of justice states that reasonable principles of justice are to be chosen under conditions of fairness. This condition is operationalized by the famous "veil of ignorance", which restricts the self-knowledge of all choosing agents. Behind the veil, individuals have general knowledge (e.g. scientific knowledge) but no knowledge of their position in society, their interests, talents, wealth and so forth. Further, the choice problem is restricted by a range of formal conditions and assumptions. Rawls' agents are for example assumed to be rational choosers, only concerned with maximizing their own utility but without the capability of assigning probabilities to outcomes. Also, the object of maximization – the currency of justice so to speak – are so-called "primary goods" (rights, liberties, opportunities, income and wealth).[4] Lastly, the candidates for principles ordering distribution of primary goods are principles common in normative social theory such as the principles of utility maximization and they are considered by the choosers in pairwise comparisons.

From a choice situation so constructed, Rawls argues that his agents, guided by the "maximin principle" of looking for the best worst outcome, would choose his two principles of justice:

> (a) "Each person has the same indefeasible claim to a fully adequate scheme of equal basic liberties, which scheme is compatible with the same scheme of liberties for all; and
>
> (b) Social and economic inequalities are to satisfy two conditions: first, they are to be attached to offices and positions open to all under conditions of fair equality of opportunity; and second, they are to be to the greatest benefit of the least-advantaged members of society (the difference principle)."
>
> (Rawls 2001: 42-43)

Getting back to the three basic elements of HCM, we can summarize that on Rawls' account the choosing agents are rational maximizers of primary goods.[5] They find themselves in the

[4] Primary goods "normally have a use whatever a person's rational plan of life. For simplicity, assume that the chief primary goods at the disposition of society are rights, liberties, and opportunities, and income and wealth." (Rawls 2001: 54)

[5] With the mentioned exception that agents cannot attach probabilities to outcomes. This is contrary to what is usually assumed by authors employing standard rational choice theory such as John Harsanyi, who we will briefly discuss in the second section of this chapter.

peculiar choice situation of having to compare and choose principles of justice behind the veil of ignorance. And they end up choosing Rawls' two principles of justice, which constitute the core normative claim of his hypothetical choice model.

The three key characteristics of HCM – hypothetical choice, deductive modeling and substantial normative claims – are quite explicit in Rawls' theory. As such, several of its components (e.g. the veil of ignorance, the problem of choosing principles of justice, Rawls' concept of rationality) obviously render the original position and its inhabiting agents hypothetical. Rawls also explicitly claims that the principles of justice follow deductively from how he has constructed his agents and their choice problem.[6] It is also clear that the two principles of justice are substantial normative claims, for they are speculations about what the agents would choose as guiding principles of basic order in a well-ordered society. This claim is, as already mentioned above, only addressed to democratic constitutional societies and not to be understood as a universal claim, in the sense of human rights. But within these limits it is clearly a claim of good social order, prescribing "the first virtues" institutions of a democratic society are to exhibit.[7] There is however an important difference between the task of deriving and of justifying such normative claims in Rawls' theory, as in HCM more generally.

Justification and Constructivism

To clarify this difference, recall that HCM aims to establish the justificatory link between claims of good social order and individual, well-considered preferences by means of a hypothetical choice situation. By presenting the original position as the paradigmatic instance of HCM, I have implied that the original position itself provides Rawls' central argument in justification of his normative claims of good social order – the two principles of justice. But this would be a misrepresentation of his theory. So ultimately, what does justify the two principles? In social contract theory the most obvious answer is: the agreement of the parties justifies whatever is agreed upon. But in a hypothetical context this answer is problematic, because there is no

[6] "The original position is also more abstract [than classic social contract theory]: the agreement must be regarded as both hypothetical and nonhistorical. (i) It is hypothetical, since we ask what the parties (as described)could, or would, agree to, not what they have agreed to. (ii) It is nonhistorical, since we do not suppose the agreement has ever, or indeed ever could actually be entered into. And even if it could, that would make no difference. The second point (ii) means that what principles the parties would agree to is to be decided by analysis. We characterize the original position by various stipulations – each with its own reasoned backing – so that the agreement that would be reached can be worked out deductively by reasoning from how the parties are situated and described, the alternatives open to them, and from what the parties count as reasons and the information available to them." (Rawls 2001: 16-17) Rawls eventually concedes that the argument from the original position does not succeed in being rigorously deductive, but is ultimately based on judgment informed and guided by reasoning. (Rawls 2001: 133-134)

[7] Recall Rawls' memorable opening statement of *A Theory of Justice* "Justice is the first virtue of social institutions, as truth is of systems of thought." (Rawls 1971: 3)

real agreement taking place and as Ronald Dworkin has pointed out early on, hypothetical agreement does not bind real citizens. (Dworkin 1975) One alternative basis for justification of the two principles would be to point to the allegedly correct construction of the choice situation and its inhabiting agents. Following this line of thought, theories of HCM aim at constructing correct normative points of view. Agreeing on justified principles is then simply a procedural result of having constructed and adopted the correct normative point of view. This does indeed seem to be the Rawlsian position. Consider the following passage:

> "I have emphasized that this original position is purely hypothetical. It is natural to ask why, if this agreement is never actually entered into, we should take any interest in these principles, moral or otherwise. The answer is that the conditions embodied in the description of the original position are ones that we do in fact accept. Or if we do not, then perhaps we can be persuaded to do so by philosophical reflection. Each aspect of the contractual situation can be given supporting grounds. Thus what we shall do is to collect together into one conception a number of conditions on principles that we are ready upon due consideration to recognize as reasonable. [...] One way to look at the idea of the original position, therefore, is to see it as an expository device which sums up the meaning of these conditions and helps us to extract their consequences."
>
> (Rawls 1971: 21)

So the original position itself is a device for deriving the correct principles of good social order, but it does not directly justify it. "It must be seen as a kind of halfway point in a larger argument, as itself the product of a deeper political theory that argues for the two principles through rather than from the contract." (Dworkin 1975: 37) This is because the original position does not provide the decisive link between the reasons of hypothetical agents and us as actual citizens. In order for us as citizens to see the argument from the original position as a justification of the resulting principles, we must be able to endorse the original position as the correct normative point of view from which to decide such matters. Thus, in constructing the original position, Rawls believes to explicate a point of view that we somehow already endorse as citizens, or would endorse after "due consideration".[8]

I will discuss this kind of constructivism in more detail below in the second section of this chapter. What we should keep in mind is that in Rawls' theory the decisive justificatory work is not done by the notion of a comprehensive agreement, but by the allegedly correct

[8]The clarification that not the choice of agents in the original position, but its correct construction from certain ideas of public political culture, does only become explicit in *Political Liberalism*. See also Peter Niesen (2016: 33).

construction of the hypothetical choice model. In this sense, HCM, according to the Rawlsian paradigm, is constructivist first and contractualist second.

2.1.3 Gaus' Deliberative Model

A more recent, comprehensive instance of HCM is found in Gerald Gaus' (2011) *The Order of Public Reason* – from now on referred to as *OPR*. I discuss Gaus' work here because his theory includes the object of my criticism in this chapter – a hypothetical choice model called "the deliberative model" – as well as what I see as a more appropriate solution to the problem of justified social order – real world norms as a uniquely justified equilibrium. I will return to the latter in following chapters. Here I only discuss Gaus' "deliberative model" as an instance of HCM.

The Deliberative Model

Gaus' overall aim in OPR is to show how a moral order, facilitating human coordination and cooperation, can be justified and internalized by a reasonably pluralistic public of free and equal citizens. One core component of his complex line of argumentation is a hypothetical choice model which he calls the "deliberative model". It is inhabited by agents called "Members of the Public". "We can understand these Members of the Public as the rationalized counterparts of real moral agents." (G. Gaus 2011: 267) They are boundedly rational beings, capable of internalizing moral rules and only act on their own sufficient reasons. In contrast to Rawls' veil of ignorance, individual beliefs, knowledge and preferences are not abstracted away in the deliberative model. Basically the idea is that Members of the Public have the same reasons we would have after having spent a respectable amount of time reflecting upon and discussing our moral convictions under ideal deliberative conditions.

Another important aspect of the deliberative model is Gaus' notion of an "optimal eligible set". He argues that the deliberative model cannot deliver a definitive answer to the justificatory problem. That is, in face of reasonable pluralism all we can hope for is to define a range of possibly justified moral rules. Defining this range is the purpose of the optimal eligible set (or OES). The baseline restrictions on the set of moral rules the Members of the Public can choose from are, firstly, a number of principles resulting from the deliberative setting and its purpose. These restrictions include requirements of reversibility, generality, publicity and Pareto efficiency.[9] Secondly and more importantly, the OES is defined as a set of proposals that "[...] consists in all those proposals that are unanimously ranked by all Members of the Public as strictly preferred to blameless liberty [...]." (G. Gaus 2011: 322) "Blameless liberty" serves as a state of nature argument, requiring that justified moral rules in the OES must

[9]For a detailed treatment of what these restrictions require see G. Gaus (2011: 294-303; 321-323).

be strictly preferred by all members of the public to a state where every individual has the blameless liberty to act according to her own standards.

In order to further restrict the optimal eligible set, Gaus goes on to abstract from all the reasons setting the Members of the Public at odds and asks: What are the basic reasons shared by all Members of the Public? His answer consists in the claim that all Members of the Public see themselves as having their own reasons and making their own decisions accordingly. Being "agents" and "autark reasoners" in this way, Gaus argues that all Members of the Public have a fundamental interest in two sets of basic rights. First, agency rights, including freedom of speech, and freedom from harm, manipulation and coercion.[10] Second, jurisdictional rights, such as privacy rights and property rights, because such rights guarantee every individual a private sphere in which she is the ultimate authority on truth, the right and the good. These two kinds of rights further restrict the optimal eligible set and constitute the core normative claims resulting from Gaus' hypothetical choice model.

Real Public Reason?

Now that we have some understanding of Gaus' hypothetical choice model, let us discuss how it instantiates the typical characteristics of HCM: hypothetical choice, deductive modeling and substantial normative claims.

Clearly the choice situation presented by the deliberative model is hypothetical in that it is not meant to describe an actual social process. As with Rawls, Gaus does not think that his model correctly describes how citizens actually think about or discuss matters of good social order. Rather the idea is again that of a construction, which shows how morality could be thought of as being justified relative to the reasons actual citizens have under conditions of ideal deliberation. The last point also marks a difference to Rawls. Gaus' members of the public are not counterfactual versions of actual citizens but merely idealizations.[11] That is, they are citizens who are assumed to have taken considerable time and effort to carefully reason about questions of morality and are ready to actually follow the guidance provided by this process of reasoning. Maybe a discussion in a seminar on moral philosophy would be an appropriate visualization of what this would actually look like. In any case, the ideal Gaus has in mind is not counterfactual in the sense that it could not or does not instantiate in reality at all. However, in spite of calling part two of OPR where he explicates the deliberative model and its conclusion "Real Public Reason", Gaus never advocates the deliberative model itself as a role model for how citizens should actually reason about moral or political matters. Thus,

[10] For a detailed elaboration of agency rights in OPR see G. Gaus (2011: 341-359).

[11] "The parties to such models are idealizations of real moral agents – they are idealized in the sense that they recognize, and judge on, their sufficient reasons in the deliberative model. They are realistic idealizations of real moral agents." (G. Gaus 2011: 266)

also Gaus' deliberative model remains a theoretical construction and in that sense entirely hypothetical.

The deductive modeling found in OPR is fairly close to the reasoning in Rawls' original position. Gaus also employs formal rational choice models in OPR, especially game theory, but his argumentative style in constructing the deliberative model is less formal. Still, the key characteristic of deductive modeling in HCM remains: The claims ultimately argued for are meant to follow directly from the specification of the hypothetical choice model. In Gaus' theory this means that the Members of the Public are claimed to all have reasons to agree on agency and jurisdictional rights, because these are the kind of things that are important to them as the practical reasoners they are.

Agency and jurisdictional rights are the substantial normative claims, resulting from the hypothetical choice model in OPR. These rights provide a fairly concrete and arguably practical view of good social order. Gaus for instance goes on to argue that socialism is not in the OES, because in light of empirical evidence, private property regimes with a high degree of economic freedom are crucial for ensuring the kind of rights and freedoms Members of the Public deeply care about. (G. Gaus 2011: 513-515) Gaus also stresses that these rights "[...] provide us with only general or abstract principles or guidelines, not with the sorts of rules that can provide the basis for firm mutual expectations about [...] what people will really do." (G. Gaus 2011: 390) The core function of agency and jurisdictional rights is to limit the set of possibly justified moral rules (the OES) which the members of the public can choose from. Thus these principles function as general guidelines for the process of choosing moral rules, which Gaus sees as something that must be thought of as an actual social process. I will pick up this line of thought in the following chapter.

Overall we can say that HCM in Gaus is a more realistic derivative of Rawls' contractualism. It is more realistic because it proposes a hypothetical choice model of ideal deliberation which, although not describing or prescribing an actual social process, could in principle take place in reality. This could happen wherever people are able and committed to adhere to high standards of rationality and reasonableness when discussing matters of good social order. However, it is not Gaus' aim to argue for actual practices of public reasoning. Rather, his substantial normative claims – the optimal eligible set constrained by agency and jurisdictional rights – follow from the idea, or model, of ideal deliberation. What actual deliberators will come up with is of no importance to this argument.

For now, we pause our inquiry into the different instantiations of HCM. I hope that the overview

over Rawls' and Gaus' versions of HCM have sufficiently exemplified the approach. In the following we will also briefly touch upon John Harsanyi's and David Gauthier's hypothetical choice models. This is of course way short of a complete overview over the use of HCM in normative social theory. Such an overview is beyond the scope of this chapter. Generally, the works of Rawls and Gaus will continue to receive the most attention in the chapters ahead. In the case of Rawls this is because his original position and constructivism constitute the most influential paradigm case of HCM. Gaus's theory on the other hand plays a more ambiguous role: HCM is clearly present in OPR, but, as we will see in the following chapters, his overall project is also a role model for the alternative to HCM I am after.

2.2 The Methodology of HCM

Above I pointed to the fact that in HCM the success of the justificatory argument relies on the correct construction of the choice problem faced by the agents, not on the idea of comprehensive agreement. Briefly and somewhat crudely put, the problem I see with this constructivism is that it is a quite arbitrary process and a source of endless debate. In more academic terms, my critique is that the kind of constructivism we find in HCM does not contain a genuine methodology. While HCM appears to make use of established methodologies, such as thought experiments or hypothetical modeling as we know it from social science, on a closer look, we find that actually HCM is just theorists constructing arguments for certain principles according to their philosophical tastes and predispositions. Thus, they end up being stuck with a problem of reasonable pluralism on the level of normative social theory that is quite similar to the original problem of reasonable pluralism on the level of citizens – i.e. the very problem HCM was meant to solve. This critical line of thought is not meant as a knockdown argument. That is, I claim to show that there are systematic reasons why HCM is always likely to fail. But I do not assert that this is necessarily the case. Hence, given that my critical argumentation is correct, it might still happen that one day some theorist presents a HCM-like theory that would strike everyone as correct. I am merely trying to show that there are systematic and inherent features of HCM that render this scenario highly unlikely.

In order to present this critical argument in more detail, we first need a better understanding of what kind of methodology or approach we are dealing with in the case of HCM. To this end, in a first step, I distinguish HCM from the well-established methods of thought experiments and hypothetical modeling as it is used in social science. Both of these methods seem intuitive candidates for explaining how HCM is done. But, as I show in the following, this intuition is clearly misleading. Thus, in a second step, I analyze HCM as a methodology in its own right. Doing so finally leads me to the formulation of my critical stance hinted at above. So let us

begin by having a closer look at the relation between HCM and thought experiments.

2.2.1 HCM vs. Thought Experiments

What is a thought experiment? Of course the nature, virtue and proper scope of thought experiments is the subject matter of many debates within philosophy and beyond. But for our purposes, a broad and fairly uncontroversial definition is sufficient:

> "Thought experiments are basically devices of the imagination. They should [...] be distinguished from counterfactual reasoning in general, as they seem to require an experimental element (i.e., visualized, touched, heard, etc.), which explains the impression that something is experienced in a thought experiment. In other words, though many call any counterfactual or hypothetical situation a thought experiment, this appears too encompassing."
>
> (J. R. Brown and Fehige 2019)

The core point I wish to make here is that HCM lacks this key feature of thought experiments: being experimental. Admittedly, the demand of being "experimental" is quite vague. So let me try to be more specific. It is commonplace in philosophy to make theoretical arguments that rest on conceptual stipulations and counterfactual premises. If we take a thought experiment to simply be an instance of such kind of reasoning where the audience is asked to follow or verify some line of thought, (to some extent) irrespective of tangible matters of fact, most of philosophy would be a thought experiment. If someone wishes to entertain so wide a conception of 'thought experiment' that is certainly fine with me. Here I am simply not interested in such a conception, because I am concerned with the experimental nature that characterizes thought experiments in a more narrow sense. This narrow conception demands an experience that is distinct from the mere entertainment of hypothetical, rational reflection. This experience is created by asking an audience to mentally enter some hypothetical scenario, which then produces the decisive individual experience. Let us consider two prominent examples.

First, think of the popular trolley problem. In this thought experiment, often used in ethics, we are asked to imagine a choice between different options of stopping or rerouting a runaway train. Typically all options lead to problematic consequences, i.e. the death of one or more individuals. Thus, what we are facing is a classic dilemma situation that does not come with a clear-cut solution but rather forces us to choose between different undesirable options. The reason why this thought experiment has been so popular in ethics and moral psychology is its usefulness in exposing our different ways of normatively thinking about the social world.[12]

[12]See for example Judith Thomson (1985) and Piercarlo Valdesolo and DeSteno (2006).

Thus, in such a dilemma scenario the interesting part is what people come up with when they enter this imaginary situation.

A second well-known example from a different field of philosophy is Thomas Nagel's bat argument. In his 1974 paper *What is it like to be a bat?* he asks his readers if they can imagine putting themselves into the mind of a bat and truly answering the question posed in the title of the paper. Nagel draws the conclusion that we cannot possibly imagine what it is really like for a bat to be a bat and uses this fact in order to make a point about the uniquely subjective character of consciousness. His entire argument is of no importance here. What matters in this context is that Nagel uses a thought experiment to prove a fact about certain limitations in our imaginative capacities. We can ourselves test Nagel's claim by asking ourselves if we can imagine what it is like to be a bat and the experience we make in doing so is what matters for the success of the argument.

The Disanalogy

There is a very intuitive connection between the kind of choice problems HCM presents and the experience of entering a thought experiment. I assume most people experience this connection, as I did, when first hearing about Rawls' famous original position and its ingenious core of making choices behind a veil of ignorance. This vivid idea seems to invite the reader to step right into an exciting thought experiment, offering a uniquely unbiased perspective on our social world. Probably any person intellectually interested in normative social matters is drawn to the attempt of looking at things from behind the veil of ignorance. But as the student of Rawls' theory of justice learns more about how the idea of the original position is constructed in detail, the first dose of disappointment kicks in. For as Rawls constructs the original position, it turns out that no one can really enter it. As I have already pointed out above, the original position is not meant to be entered into by us, but is already occupied by identical, formally defined agents. These agents are not individual choosers, at least in the sense that we think of ourselves as choosers, with personal goals, experiences, an individual attitude toward risk and so on.[13] Thus, what started out as a lively and inviting thought experiment is really just a formal argument, presenting premises and drawing conclusions.[14]

More generally speaking, in HCM the choice problem is not constructed in order to be entered

[13] "So although the original position begins by posing a problem of collective choice, the problem is reduced to the Kantian problem of public legislation by one person." (G. Gaus 2011: 38)

[14] In *A Theory of Justice* and *Political Liberalism* Rawls does not mention the term thought experiment at all. In the 2001 restatement Rawls uses the term loosely, without implying that he is referring to the conception of 'thought experiment' I make use of here (Rawls 2001: 17,83), whereby he does continue to stress the deductive logic of his argument (Rawls 1971: 104-105, 2001: 16-17, 82)). Nevertheless, as in Michael Lessnoff (1986: 159) it seems commonplace to refer to Rawls' hypothetical choice model as a thought experiment, simply because it contains a counterfactual situation.

into imaginatively as is the case with genuine thought experiments. There is really nothing experimental about the hypothetical choice at all. As 'modeling' suggests, HCM is rather about specifying a choice problem to the degree that one can rationally deduce the correct choice from the setup of the scenario.
Putting aside Rawls theory, it is of course not a necessary feature of HCM that the model cannot be imaginatively entered into. I could for instance imagine being a Member of the Public in Gaus' deliberative model. However, and this is the key point, the argument from the deliberative model depends in no way on my experience when trying to do so. What renders the argument correct or false is not experience, but reason. This *deductive*, in contrast to an *experimental* approach is no mistake or accident on the part of the theorist. In fact, I am quite certain that Rawls and Gaus (just as the other theorists we turn to below) would be happy to admit that their argumentation is deductive and not experimental in this way. Thus, showing that models of hypothetical choice are not thought experiments should neither be controversial, nor is it meant as an objection against any theory employing HCM. My aim here is merely to separate the two notions and to show that HCM cannot draw on the kind of experimental arm-chair evidence that thought experiments might provide.

2.2.2 HCM vs. Modeling in Social Science

Now that we have clarified that the *hypothetical* nature of HCM does not imply the presence of a thought experiment in the narrow sense, we turn to the method of *modeling* in HCM and distinguish it from modeling in social science. As with thought experiments and HCM, there is an obvious connection between HCM and modeling in social science. Namely, that models in economics or political science are often highly abstract to the degree that they are also somewhat counterfactual. Further, they typically apply a deductive logic in order to explain or predict what some actor does, based solely on what follows from the set up of the model. But there are also important differences. The main one being that models in social science model actual social processes whereas HCM does not.

Modeling in HCM

In *A Theory of Justice* Rawls only speaks of modeling on two occasions (Rawls 1971: 112, 165). In *Political Liberalism* as well as in his 2001 restatement, he uses the concept of modeling extensively, specifically for describing what the device of the original position and its components do. Consider for example the following passages:

> "Keep in mind throughout that, as a device of representation, the original position models two things. First, it models what we regard – here and now – as fair

> conditions under which the representatives of citizens, viewed solely as free and equal persons, are to agree to the fair terms of social cooperation (as expressed by principles of justice) whereby the basic structure is to be regulated. Second, it models what we regard – here and now – as acceptable restrictions on the reasons on the basis of which the parties (as citizens' representatives), situated in those fair conditions, may properly put forward certain principles of justice and reject others."
>
> (Rawls 1993: 80)

Rawls then goes on to explicitly point to the similarity between how modeling is employed in his theory and in social science:

> "Note first the similarity between the argument from the original position and arguments in economics and social theory. The elementary theory of the consumer (the household) contains many examples of the latter. In each case we have rational persons (or agents) making decisions, or arriving at agreements, subject to certain conditions. From these persons' knowledge and beliefs, their desires and interests, and the alternatives they face, as well as the likely consequences they expect from adopting each alternative, we can figure out what they will decide, or agree to, unless they make a mistake in reasoning or otherwise fail to act sensibly. If the main elements at work can be modeled by mathematical assumptions, it may be possible to prove what they will do, ceteris paribus."
>
> (Rawls 1993: 81)

Although the assumptions differ greatly, something similar could be said about Gaus' deliberative model. Both theorists use rational choice style modeling in normative theorizing and at least at first sight this seems to be very similar to how this is done in social science. In fact, in their 2017 revision of the Stanford Encyclopedia of Philosophy article on *Contemporary Approaches to the Social Contract*, Fred D'Agostino, Gerald Gaus and John Thrasher present modeling as a core characteristic of contemporary social contract theory in general. And they also provide some brief remarks on the nature of such modeling:

> At the simplest level, models take something complex and make it simpler. [...] Models involve abstraction and idealization, but they do more than that [...]. Modeling seeks to isolate the important features of the target phenomena, allowing the modeler to understand and manipulate important elements of the phenomena in

> simulations. John Rawls's representatives to the original position, for instance, are not only abstractions of real persons. They are idealizations that isolate particular aspects of persons that are relevant to justification as a choice, specifically their thin theory of rationality, and their values (in the form of primary goods). Isolating these features is important for modeling the agreement procedure in Rawls's theory. The social contract models our reasons for endorsing and complying with some set of social rules or institutions. How the theory does this depends on the assumptions made and the specification of the parameters.
>
> (D'Agostino, G. Gaus, and Thrasher 2017: 80)

These brief remarks on the nature of modeling in normative theory clarify that models in HCM are more than an abstraction or mere simplification of a complex phenomenon. They are manipulations, or, to use an expression more common in philosophy that we have already come across above, a "construction" of something in a way that "isolates" some features rather than others, because this is the most appropriate way to proceed according to some theorist. Two obvious questions resulting from these reflections are one, *What is the object of modeling in HCM?* and two, *How do we know if modeling has been done correctly?*. Both questions do not possess an obvious answer in HCM, but they do in social science. Which brings us to the decisive differences between modeling in HCM and disciplines such as economics and political science.

Modeling in Social Science

Rational choice modeling in social science usually works as an "as-if explanation". They model actual social processes by means of abstraction, idealization and formalization. That is, a complex phenomenon is boiled down to a more simple depiction, where also some things may be assumed to be more ideal (e.g. the existence of perfect information) than they actually are. Further, this simplified and idealized picture is translated into a formal framework – utility functions, game theory, bargaining theory or the like – which allows the authors to be precise about the assumptions they make and the conclusion that follows from them. Essentially, the model produces an explanation that is itself somewhat counterfactual and not necessarily a true representation of the phenomena or process it is trying to model. It is rather an *as-if* formalization of what actually happens. The point of the exercise being the provision of verifiable predictions or the identification of general underlying logics or mechanisms of the modeled process.[15]

[15]Perhaps *the* classic explication and defense of this methodology is Milton Friedman's (1953) essay on *The Methodology of Positive Economics*. He claims that it is entirely beside the point how realistic the assumptions

To the end of having a prominent example in mind that was familiar to Rawls' and is cited by him in *A Theory of Justice*, let us consider Anthony Downs' (1957) *An Economic Theory of Political Action in Democracy*. Downs' aim is to show how democratic government can be modeled by means of economic rational choice theory. He paints a picture of citizens as voters on the one side, who vote for certain parties (and thereby the policy or ideologies they stand for) in order to maximize their expected individual payoff. On the other side are political parties promoting policies in order to maximize votes and be elected for office. With this highly simplified picture in mind, Downs discusses a whole range of issues, such as the crucial role of knowledge, lobbyists, and the dynamics between voters and parties. One of his main results is that the distribution of voters in the policy space is decisive for what kind of party system (i.e. a two or more party system) prevails and what the dynamics of that system are. Besides being a classic and highly interesting read, Downs' model of democratic politics nicely illustrates how modeling in social science is usually done. Here I want to stress three characteristics. First, the object of modeling is a fairly specific social process – in this case the political dynamics in a constitutional democracy such as the United States. Second, the appropriateness of simplifying assumptions is discussed in relation to the actual social process in question and the scientific aims of the model. Downs for instance assumes that all agents are rational in the economic sense (they maximize individual payoff), that the policy space is a one-dimensional continuum between extreme right and extreme left and that the majority winner gains "ultimate" political power. None of these assumptions are necessarily true in political reality. Thus, right from the start, it is clear that also the conclusions drawn from such a model will not be an exact representation of reality. Rather, the question is whether the assumptions made are a simplification, close enough to reality to allow us to draw partly true conclusions about the more complex reality of the modeled process. Downs himself engages in the discussion about the proper conception of individual rationality in the political realm in a later paper. (Downs 1962) Third, the success of the model and its conclusions can also be assessed relative to social reality. For instance, the probably most well-known hypothesis of the model proposed by Downs and his predecessors is the median voter theorem.[16] This denotes the claim that given certain more specific conditions,[17] parties will converge on the preferences of the median voter. Although the assumptions necessary to support this claim are usually violated in several respects in political reality, it still offers an interesting explanation

are, as long as the model delivers valid predictions. There is of course a vast literature about such claims, the nature of modeling in social science more generally and the different kinds of models there are and the functions they serve. For a brief and recent summary of hypothetical models in social science see Chiara et al. (2017).

[16] Downs' work explicitly builds on the earlier analysis by Harold Hotelling (1929) and arguably also on Duncan Black (1948).

[17] Most importantly a normal curve distribution of voters in the one-dimensional policy space.

for the phenomenon that in political systems with two large popular parties, these parties sometimes have a tendency of convergence between platforms or proposed policies.[18] Although neither the assumptions nor the outcomes are comprehensive truths, they could very well be an approximation of the truth, a piece of the puzzle so to speak, in understanding a complex social process, difficult to grasp in total by the means of social science.

The Disanalogy

In a nutshell, models such as Anthony Downs' economic model of politics are simplified, formalized and somewhat counterfactual representations of actual social processes. They rebuild a piece of the social world in theory *as if* it were more simple and clearly structured than it actually is. If and to what degree such a model tells us something interesting about the world we actually live in is to be assessed by comparing the real social process and the in- and outputs of the model (i.e. its assumptions and resulting explanations or hypotheses). With this in mind, we can now see an obvious difference between hypothetical modeling in social science and HCM: If we focus on the agreement reached within the hypothetical model, HCM clearly does not model any actual social process. For this is precisely what the social contract tradition has done away with since the hypothetical turn. Hypothetical models in social science are counterfactual in that they construct a somewhat incorrect description of an actual process. HCM on the other hand is a *purely* counterfactual fabrication: The agreement, deliberation, bargaining or choosing "modeled" in HCM does not take place in reality at all. D'Agostino, G. Gaus, and Thrasher (2017) thus speak of a "doubly" hypothetical model about what would be chosen in a choice that never materializes in actual society.

From the point of view of hypothetical modeling in social science, this is an odd methodological framework. For what makes a model meaningful and productive there, is its relation to the actual social process being modeled. The model itself is a theoretic makeshift device, floating in the realm of thought and fiction. What grounds this fiction are its two points of intersection with reality: the construction of the model and its conclusions, which should at least in principle be testable hypotheses. In HCM there is no such systematic relation between the model and the social process being modeled. Such models are free-floating theoretical fictions with a relation to reality that is difficult to grasp.

One might respond that this is a natural and necessary byproduct of professing *normative* social theory. Because in normative theory we are not primarily trying to model how the world is, but how it should or would be under more ideal conditions. This is a fair point. But it does not answer the question of whether HCM actually contains a cogent methodology. Here my aim is to show how modeling in social science and HCM differ: In social science, models

[18] For a fairly recent and detailed discussion of the convergence thesis see Bernard Grofman (2004).

usually model actual social processes and can be tested relative to this modeled reality. In HCM the modeled agreement does not seem to correspond to an actual social process.

2.2.3 HCM as an Approach in its Own Right

So far we have established that HCM is neither a thought experiment, nor a case of hypothetical modeling as we know it from social science. Now, one critique, directly building on these insights, would state that HCM is really a misguided combination of the two methodologies: On the one hand, it takes the purely hypothetical nature of thought experiments and thus excludes the modeling of actual social processes. On the other hand it employs the deductive logic of formal modeling and thus excludes the experimental aspects of thought experiments. More simply put, HCM appears to combine two methodologies by leaving out two of their necessary features. As if one were to combine a cooking pot and a kitchen blender by only taking the lid of the pot and the jar of the blender, being left with something of which it is at the least unclear how it could be used to prepare any food. This line of critique might not be far from the truth, but it would be unfair to leave it at that. Rather, before reaching any critical conclusion, we should look at HCM as an approach in its own right, in order to investigate whether it contains a promising alternative to the two methods discussed so far.

Constructivism in HCM

If we are to consider the methodological nature of HCM within the Rawslian paradigm, we have to return to the notion of *constructivism*. Constructivism has become somewhat of a hot topic in political theory, moral theory and metaethics.[19] The important thing to keep in mind in this chapter is that I am not concerned with a critique of constructivism in general, but only of constructivism in HCM. In fact, and as we will see in the following two chapters, I come to term my own alternative to HCM a kind of constructivism – namely "Embedded Constructivism".
With this in mind, let us turn to Rawls' first introduction of constructivism in the *Dewey Lectures*:

> "What justifies a conception of justice is not its being true to an order antecedent to and given to us, but its congruence with our deeper understanding of ourselves and our aspirations, and our realization that, given our history and the traditions embedded in our public life, it is the most reasonable doctrine for us. [...] Kantian *constructivism holds that moral objectivity is to be understood in terms of a suitably*

[19]For a fairly recent overview of the different strands and issues in the debates on constructivism see James Lenman and Yonatan Shemmer (2012).

> *constructed social point of view that all can accept.* [...] Whether certain facts are to be recognized as reasons of right and justice, or how much they are to count, can be ascertained only from within the constructive procedure, that is, from the undertakings of rational agents of construction when suitably represented as free and equal moral persons."
>
> (Rawls 1980: 519), my italics

According to Rawls' conception of justice as fairness, the "suitably constructed social point of view" is his original position.[20] Later on, in his *Political Liberalism*, Rawls defends a "political constructivism", holding that only the principles of justice are constructed, whereas the procedure for arriving at them is "simply laid out". (Rawls 1993: 103-104) To my mind this conceptual variation in *Political Liberalism* is implausible. In Rawls' theory, as in HCM more generally, the principles are *deduced* or at least *derived* from the hypothetical choice situation specified by the theorist. To say that they are "constructed" does not add anything. Rather, it is the complicated matter of setting up the correct choice situation – the "suitably constructed social point of view" – that amounts to a construction. Thus, in my view, if the term is to mean anything distinct, 'construction' should refer to the art of generating suitable social points of view.[21] In any case, this is not a substantial objection but merely a conceptual discrepancy.

So, given that constructivism in HCM consists in constructing "suitable social points of view" for deriving normative principles, how does it work? This is a difficult question to answer. Nevertheless, there are at least three aspects we can point to here. First, every construction needs some building material. Thus, Rawls stipulates certain elements or ideas which form the building blocks of the construction. Such ideas include the idea of fairness, of practical rationality, of different principles of justice, of a well-ordered society and of a person in such a society. These ideas can be found in "public political culture as well as in citizens' shared principles and conceptions of practical reason."(Rawls 1993: 93) So the building material itself are ideas or convictions citizens in western democracies allegedly already endorse or would endorse on due reflection. After having identified such ideas, the task of construction seems to specify and situate them in a way that leaves us with a *conclusive* choice problem of choosing principles of good social order.

[20] I consider the notions of a 'social point of view' and that of a 'hypothetical choice model' interchangeable here, although one could of course think of instances of the former that are not an instance of the latter.

[21] In a different passage of *Political Liberalism* it seems that Rawls also holds this view: "The initial situation is an attempt to represent and to unify the formal and general elements of our moral thought in *a manageable and vivid construction* in order to use these elements to determine which first principles of justice are the most reasonable." (Rawls 1993: 275, my italics)

Second, the choice problem so constructed offers a *theoretical* procedure for arriving at justified principles of good social order. This procedure is not a social process, but an exercise of reason on part of the theorist, which is why Stephen Darwall et al. (1992) refer to a "hypothetical proceduralism".[22]

Third, note that within the Rawlsian paradigm, there is a test for the reasonableness of any construction of principles of good social order. This test has to do with the notion of *reflective equilibrium* and of an *overlapping consensus*. Simply put, Rawls' hypothesis is that justice as fairness constitutes the most coherent view of the problem of choosing good principles of basic social order for all citizens, if they would actually take the time to think things through. More precisely, Rawls' hope is that his construction explicates an "overlapping consensus" between all reasonable citizens[23], whereby they should see this if they employ their powers of reason to reach "general and wide reflective equilibrium".[24] That is, if they bring their individual, well-considered judgments on matters of normative theory, normative principles and judgments in line with all such well-considered judgments others might have.

Let me sum up what we have learned so far about constructivism in HCM. Constructivism in HCM means excavating certain ideas from a society's political culture, tradition and history of thought, and certain general facts about social life in cooperative societies. These ingredients are then specified and arranged in a way to provide a social point of view – a hypothetical choice situation – in order to derive normative principles of good social order.

This sounds like a terribly complicated task and a source for many, differently constructed social points of view. Which is exactly what we see in the literature discussed in this chapter, where different theorists present us with different kinds of hypothetical choice situations. So what then constitutes standards of correctness that we can apply to choice situations and their outcomes?

As pointed to above, in social science scholars may come up with different models for the same social process which can then – at least in principle – be tested and compared with regard

[22] "Construction then enters at two points: the theorist constructs a social point of view, a hypothetical circumstance for the choice of moral principles, and hypothetical choosers construct the moral principles that best serve their ends. The hypothetical choosers are "agents of construction" in both senses: the theorist constructs them and they construct principles." (Darwall, Gibbard, and Railton 1992: 139)

[23] "As a political conception it [justice as fairness] aims to be the focus of an overlapping consensus. That is, the view as a whole hopes to articulate a public basis of justification for the basic structure of a constitutional regime working from fundamental intuitive ideas implicit in the public political culture, and abstracting from comprehensive religious, philosophical and moral doctrines. It seeks common ground – or, if one prefers, neutral ground – given the fact of pluralism. This common ground is the political conception itself as the focus of an overlapping consensus." (Rawls 1993: 192)

[24] "That is its [i.e. the theory's] primary aim: to be presented to and understood by the audience in civil society for its citizens to consider. The overall criterion of the reasonable is general and wide reflective equilibrium [...]." (Rawls 1993: 384)

to how well they explain or predict the targeted social process. Therefore, the correctness of the model itself is of secondary importance. In HCM, however, there is no obvious relation between the model and some underlying social process. So is there any systematic relation between theoretic construction (or modeling) and social reality in HCM? I do not believe that there is one clear answer to this question. Rather, I attempt to show in the following that there are at least two plausible readings of how we can understand the core methodological nature of HCM.

The Normative and the Empirical Reading

The two different readings concern the overall status of the model and can be summarized as a distinction between a model of how we *should* think about the problem of justified social order and *re*constructions that explicate how we *do* think about social order. This distinction is analogous to the difference between descriptive and normative decision theory. In descriptive decision theory the modeling of some situation is about predicting what rational agents will do. Normative decision theory on the other hand is about what rational people should do. The difference between the two approaches stems from the different stance toward the idea of rational agency. This idea is what drives both kinds of analysis, but in descriptive theory it comes in the form of an assumption about what people are already like, whereas in normative theory it takes the form of an ideal people should live up to. Thus, in the case of descriptive modeling, if the model fails to predict the actions of some person, the fault is with the fit of the model, not with the person. In the case of normative modeling, things are the other way around. Someone failing to act according to the normative model does not pose a challenge to the model, but may be criticized for failing to act rationally.

In HCM the overall status of the choice model can similarly be interpreted in a descriptive and in a normative way. Consequently, there are at least two possible readings of what constructivism in HCM is all about.[25] Thus my aim now is to lay out both views in order to carve out the different consequences and problems they imply.

The Normative Reading

According to what I call the normative reading, models of hypothetical choice are primarily models of how we *should* think about the problem of justified social order. The perspective or

[25] Daniel Gaus (2013) has pointed to a normative and an empirical interpretation of Habermas' method of rational reconstruction. However, his take on the empirical perspective is distinct from what I call the empirical reading in that it he stresses an alleged explanatory role of reconstruction in Habermas. Whereas the empirical reading that I refer to in the following has an explicatory rather than an explanatory role. Further, Daniel Gaus relies on a distinction between constructive and reconstructivist theories of justice, which I do not see in the literature. At least not if this distinction is meant to point to an important difference between Rawls and other authors.

social point of view they provide are philosophical innovations driven by philosophical ideas that normal citizens do not necessarily have.
Thus, this viewpoint does not model how citizens necessarily do or ever will think about the matter. The model and its outcomes rather explicate how they *should* think about it. Understood in this way, the construction produces a normative meta perspective, which can only be taken, discussed, understood and criticized from within normative social theory.

In OPR, the normative reading is apparent in Gaus' reflections on how we can gain insights on "true morality" – by constructing the right "moral point of view on morality". In this passage, Gaus informs us that it is not enough to have an account of the different "positive moralities" we actually have in order to gain a critical perspective on social order. What we further need is a conception of "true morality" that brings with it the right social point of view for evaluating positive morality. This move seems to introduce the kind of extra normative level, above and beyond existing normative thought, that is typical for the normative reading.[26]
The normative reading is also exemplified by Gaus' testing conception of public reason. Thereby the idea is that we can take his hypothetical choice model and its outcomes (the idea of an optimal eligible set constraint by agency and jurisdictional rights) as a means of testing the justifiedness of existing or proposed institutions. Thereby the test is a one-way critical perspective on social arrangements – Gaus never considers the possibility of his theory itself being tested by actual societal deliberations.
In Rawls, the normative reading becomes apparent if we restrict our view to the idea of the original position that produces principles of justice as the first virtues of social institutions. As I also pointed to above in the disanalogy between HCM and thought experiments, the agents and their reasoning in the original position fundamentally differ from how we as normal citizens reason and choose. Further, Rawls presents justice and choosing principles of justice as *the* central normative perspective on social order. With this he creates a separate normative sphere – the sphere of justice – beyond things like laws, constitutions and political intuitions that are familiar to citizens.

The main methodological implication of the normative reading is that everything hinges on the correct construction of the model. This is because it is irrelevant whether the model itself and its outcomes reflect what we do or will think as citizens. The model does not have any empirical implications. Once it is presented and the resulting principles are derived, there is no further test in respect to these outcomes. This lays a heavy burden of justification on any theorist of HCM.[27] For she has to somehow show, why, in virtue of its construction, her model

[26] See G. Gaus (2011: III.10.4) and my more extensive treatment of this passage in *Subsection 3.1.1.*
[27] Jürgen Habermas (1995: 118) also thinks that in Rawls' constructivism the theorist carries a heavy burden of

is correct.[28]

The Empirical Reading

According to what I call the empirical reading, models of hypothetical choice are essentially models of how we think about the problem of justified social order. In other words, often used in the public reason literature, they model "the reasons we have", whereby "[t]he aim is to model the reasons of citizens, and so we ask what they would agree to under conditions in which their agreements would be expected to track their reasons." (D'Agostino, G. Gaus, and Thrasher 2017)

On this account, modeling does not construct a unique, theoretic point of view. Instead it *re*constructs a point of view that is already – somehow – present in existing thought and practice. It explicates what we as actual citizens think or endorse. We may not necessarily have or understand these reasons explicitly – perhaps they have not been carved out in a coherent and systematic manner yet – but we could be enlightened to see that they follow from how we think about matters of good social order if we would take the time to think things through (collectively).

To clarify, I understand 'construction' as the general activity of producing normative, theoretic models, such as hypothetical choice models. 'Reconstruction', on the other hand, I take to be one aspect of this construction that some constructivists tend to highlight: a rationalized explication of things that already exist in normative thought and practice.[29] If this is done successfully, even modeling according to the empirical reading can offer a critical perspective on what citizens actually think and what institutions they establish. The two main differences in respect to critical outcomes of the model and in contrast to the normative reading are, firstly, that according to the empirical reading the provided criticism is a pointer to inconsistencies in, or underappreciated implications of existing normative thought.[30] Whereas according to the normative reading, it is an independent critical perspective on existing normative thought and practice. Secondly, in contrast to the normative reading, the critical perspective is not a one-way relation according to the empirical reading. Here, thoughts and institutions found in actual societies *can* challenge the conclusions of the model. For if it turns out that we as actual citizens reject the model's conclusions even on due reflection, this would pose a serious challenge for a model of the reasons we have.

---

justification. Although I am doubtful whether he can successfully jettison this burden himself. On this see my discussion of Habermas' constructivism in *Subsection 3.3.1.*

[28] "Judgments are reasonable and sound if they result in following the correct procedure correctly and rely only on true premises." (Rawls 1993: 102)

[29] For an example also see my discussion of Jürgen Habermas' constructivism in *Subsection 2.3.3.*

[30] As a rejoinder to critical theory we might speak of an immanent critique here.

In Gaus, the empirical reading is exemplified by the rather realistic conception of the ideal reasoners in his hypothetical choice model. These surrogates of real citizens are assumed to be endowed with sufficient intellectual capacities, which they put to use under ideal deliberative conditions. Besides these idealizations, they basically reason as we do as citizens. Thus it seems plausible to assume that "[...] your surrogate tracks your reasons – the reasons *you have*." (G. Gaus 2011: 265) Gaus also stresses that any successful argument from abstraction must fulfil a requirement of full justification according to which the conclusions reached on the abstract level are confirmed: "When the abstraction is lifted, and the deliberators are aware of the full range of their evaluative standards, the conclusion reached via abstraction must not be overturned." (G. Gaus 2011: 335-336)
In Rawls, the empirical reading emerges if we take a broader look at Rawls' theory, especially after the Dewey Lectures and his reply to Habermas. Here Rawls becomes more concerned with the stability of his conception of justice in pluralistic societies so that it can fulfill its practical function of presenting "itself as a conception of justice that may be shared by citizens as a basis of a reasoned, informed, and willing political agreement." (Rawls 1993: 9). In order to show that his conception of justice can be stable, Rawls introduces the idea of "public justification". Public justification is achieved when citizens ensure each other that the conception of justice in question is fully justified to them and that they have shared ("political") reasons for endorsing it. (Rawls 1993: 367)
Taking these passages at face value, we may conclude that political constructivism is empirical modeling in that, one, from the perspective of constructing the model, it models how we as citizens of western democracies think about the good of democracy. Two, from the perspective of drawing conclusions, the model carves out abstract conceptions of good social order that will be affirmed by citizens if properly considered.

The main methodological implication of the empirical reading is that the normative modeling is closer to modeling in social science. Normative modeling, then, also has some actual social process as its object: Normative reasoning of citizens as it would actually take place under ideal conditions. Consequently, we can interpret the hypothetical choice model and its outcomes as a hypothesis of how we as citizens actually think – or what is implied by what they think and value. This hypothesis is testable in principle. That is, it may be very difficult to actually test it in an existing society. But we can at least derive what a physically possible test would look like. Say, a long-lasting deliberative process between all citizens under ideal conditions for such an event. Thus the correctness of a normative model is eventually to be tested externally and empirically. That is, while internal coherence is of course still important, a correct model produces outcomes that are eventually excepted by actual citizens on due consideration.

2.3 Two Criticisms of HCM

This brings me to the formulation of my actual criticisms of HCM. As the attentive reader might have anticipated, I offer two lines of criticism, corresponding to the two different readings of the methodological nature of HCM. If we take the perspective of the normative reading, the problem is this: Constructivism in HCM, according to the normative reading, commits the theorist to the construction of *correct* hypothetical choice models. At the same time, there is vast disagreement between theorists regarding the appropriate building material, the appropriate arrangements of such materials as well as the purpose of construction. Further, constructivism in HCM does not provide us with any helpful standards or tools, let alone methods for resolving these differences. Even worse, it invites theorists to reproduce their different philosophical tastes and predispositions. Thus, constructivism in HCM reproduces the pluralism we already have on the level of normative social theory instead of being helpful in arriving at outcomes that have a reasonable chance of being accepted as correct.
If we take the perspective of the empirical reading, the problem is this: HCM, according to the empirical reading, includes an illegitimate abstraction in that it abstracts from the very thing that it is meant to model: the reasons we have. This is because HCM, in order to be successful, has to construct a conclusive choice problem, which remains a theoretical fiction in light of the reasons we have in social reality. Thus I conclude that HCM is generally the wrong approach for modeling the reasons we have.
In the following I consider both lines of critique successively in detail.

2.3.1 One: Lost in Pluralism

The observation that motivates the first line of criticism is expressed casually by a pointed remark of Brian Skyrms.

> "Traditional theory asks what kind of contract would have been reached by rational, reasonable agents if they were in the position of setting up an ideal contract. [...] Skeptics, since ancient times, have pointed out that different cultures have arrived at different social contracts. This is dismissed as saying that some, or perhaps all, are not rational or not reasonable. A contemporary skeptic might point out that one leading theorist at a leading institution of higher learning may arrive at one contract, while another leading theorist with an office down the hall might arrive at another, while each maintains that any rational reasonable person would agree with his view.
>
> (Skyrms 2016: 1089)

From within the circle of public reason theorists Fred D'Agostino states:

> "[T]he project of public reason starts, in diversity, [...] this empirical diversity must be accepted, if not as a given, then anyway as an unavoidable starting point for our attempts to identify the lineaments of a social order that can pass appropriate tests of its legitimacy and, accordingly, of the normative hold on people of its deliverances. But all this comes unstuck, of course, if the diversity at this "ground-floor" level, of the kinds of values that underpin concrete choices made by particular individuals in ordinary social settings, were simply reproduced at a more abstract theoretical level—at the level, say, of the standards which, as I put it, underpin the idea of public reason, the very idea that is being wheeled out, anyway by Rawls, Gauthier, and others, to address the issue of legitimacy and normativity in the face of [...] evaluative diversity."
>
> (D'Agostino 2013: 130)

This is indeed a fitting summary of one of the more problematic symptoms of HCM: Divergent models and divergent outcomes. But I think the underlying problem deserves a more systematic treatment. Let us proceed by having a look at the dispute between John Rawls and John Harsanyi, for it provides us with an excellent example of the problem with pluralism on the level of normative social theory.

The Rawls-Harsanyi Dispute

John Harsanyi presented something close to the original position several years before Rawls' publication of *A Theory of Justice* and he continued to revise and restate his argument over the years.[31] Harsanyi's hypothetical choice model is the "equiprobability model". This choice model is also inhabited by rational choosers, who are assumed to be behind a thin veil of ignorance. Behind it, the choosing agents know pretty much everything there is to know about their society besides the crucial information of which citizen they actually are. Agents are further assumed to attach equal probability to being any of the actual citizens under consideration. Given these constraints, Harsanyi uses normative decision theory to model a choice situation analogous to a choice under uncertainty in standard decision theory and concludes that agents will choose the utilitarian principle of maximizing average expected utility as the basic guiding principle of justice.

[31] For the first statement of his basic argument see John Harsanyi (1953, 1955). For a more recent and comprehensive account see especially Harsanyi (1978).

Michael Moehler (2015) offers an excellent analysis of the discrepancies between Rawls' and Harsanyi's accounts. He argues that in spite of what many commentators have said, the Rawls-Harsanyi dispute is not primarily about the correct application of normative decision theory, but about a difference in the "moral" assumption both authors bring to the table. More precisely, Moehler shows that Harsanyi's and Rawls' hypothetical choice situations both model the ideal of impartiality by denying the choosing agents the knowledge of who they actually are. However, the ideal of equality is modeled differently in both accounts. Harsanyi models equality in a distinctively utilitarian sense, which sees individual citizens as utility functions and choosing agents as maximizers of expected utility, for whom utility created by different citizens is equally important and comparable. Moehler summarizes this feature by saying that besides the ideal of impartiality and equality, Harsanyi also models an ideal of *impersonality*. Rawls anti-utilitarian account on the other hand rejects this impersonal perspective. By withholding more knowledge from the choosing agents, not allowing them to attach probabilities to outcomes and by making their choice about primary goods, he models a different ideal of equality, stressing the ideal of autonomy and the separateness of persons. Moehler concludes that

> "[a]s such, there is no winner in the Rawls–Harsanyi dispute. Instead, the dispute merely clarifies the moral ideals and their formal representations that need to be assumed in order to justify either Rawls' contractualist principles of justice or the average utility principle."
>
> (Moehler 2015: 3)

> "In this sense, Rawls' original position and Harsanyi's equiprobability model represent only two possible moral decision situations. Many other moral decision situations are conceivable."
>
> (Moehler 2015: 15)

And more generally Moehler maintains,

> "[...] different moral decision situations model different moral ideals and, consequently, may justify different conclusions about justice. Thus understood, the Rawls–Harsanyi dispute offers a promising starting point for future research that can deepen and enrich our understanding of the demands of justice."[32]
>
> (Moehler 2015: 3)

[32]Following his own advice, Michael Moehler (2018) has published a contractarian HCM-style theory.

I do very much agree with Moehler's analysis, but I reject his general conclusion regarding the promising nature of further theories employing HCM. As I see it, Moehler has precisely shown how different theorists with different (philosophical) convictions and preferences tend to construct different hypothetical choice models, leading to different principles of good social order. In his analysis, Moehler focuses on the different "moral" assumptions theorists bring to the table. But there is more to pluralism on the level of normative social theory. Let me be more specific.

Pluralism in Normative Social Theory

Pluralism in normative theory with respect to HCM has three dimensions. The first dimension results from the different views of the basic problem of justification. All hypothetical choice models seem to address some problem of choosing meta principles of good social order, but these problems are framed in different ways. They are for instance framed by broad, theoretic concepts such as 'justice' and 'morality', while there is little agreement on the meaning of such terms and how they are properly related. The second dimension of pluralism results from how different theorists pick out different building materials – e.g. different ideas, facts and ideals – as relevant and appropriate for the construction. Different thinkers view the social world differently. Consider for example Rawls' idea of a well-ordered society, which plays a central role in his theory but is absent in other theories. The third dimension of pluralism concerns the way in which the construction is done: How does one model the raw building material into a conclusive decision problem?

Before we discuss the different dimensions in turn, let us briefly reflect on the nature of of pluralism in normative social theory. Is it perhaps a superficial kind of disagreement that will eventually be overcome, or is it a case of *reasonable* pluralism? Essentially the question is whether we believe the disagreement to be due to ignorance, mistake or insufficient reasoning, or due to the fact that this is a matter where even the most reasonable people can always have different opinions. My position is that pluralism in normative social theory with respect to HCM is of the latter kind. But I cannot prove that this is correct. In general it is difficult to settle whether some case of pluralism is of a superficial or of a reasonably persistent nature. For as individuals in a divers group we often tend to think that we are right and that the others are just not getting it. And there is of course no way of knowing for sure whether some case of pluralism will not eventually prove superficial. To judge that some case of pluralism is reasonably persistent is an inductive inference, drawn from the experience of respective debates up till now. Therefore my strategy here is to firstly acknowledge that I might be wrong. That is, it might turn out that the pluralism in question is in fact superficial and that consequently one instance of HCM will eventually prove correct, at least with respect to a specific problem

of justifying social order. Secondly, I back my thesis of reasonable pluralism in normative social theory inductively by pointing to the fact of persistent, reasonable disagreement we have actually observed so far. Further, there is the problem that within HCM there are no tools available to make any progress in face of persistent pluralism in normative social theory. This should lend significant support to my claim that reasonable pluralism in normative social theory, at least in respect to HCM, is of a persistent nature.

The first dimension of pluralism seems to be the least problematic. For the difference in conceptions of the basic problem of justification could simply be due to theorists being concerned with different objects of justification. So for instance, one theorist may be concerned with formal and the other with informal instances of social order. Thus they use different basic concepts and develop different lines of argumentation. With respect to HCM and the authors we have discussed so far, however, this is not the case. All of them aim at the justification of basic principles of social order while conceptualizing this target in different ways: Rawls speaks of justice and ends up with distributional principles, while Gaus speaks of morality and ends up with a set of abstract rights. These authors have a fairly similar object of justification, while entertaining quite different conceptions of it. Further, there is no agreement about the correct conception of morality or justice in sight. Thus Wilfried Hinsch notes with respect to 'justice':

> "The uncomfortable truth, however, is that well-informed people who are quite willing to live up to the demands of justice and who believe in fair reciprocity also often disagree about what basic justice requires. Justice is a notoriously contested notion, not only in politics but also in moral philosophy."
>
> (Hinsch 2018: 103)

The same holds true for 'morality'. Generally speaking, the problem here is that theorists begin with theoretic, normative conceptions of social order, while such conceptions are one of the things different theorists typically disagree about. Essentially, conceptions of morality and justice remain contested and illusive in normative social theory and I am tempted to conclude that they are useless in finding a common understanding of the problem of publicly justified social order. The remedy for this problem I propose in the following chapter consists in starting with an empirical and descriptive account of social order.

The second dimension of pluralism is that of differences in the relevant building material for constructing hypothetical choice models. Now, given that, as we have just seen, theorists start from different conceptions of the basic problem of justification, it is of course understandable that they point to different ideas, assumptions and ideals as their building material. But

again, looking at the proposals of just one theorist, there seems to be a deeper problem of conceptualization here. Rawls for instance might have hoped that his core conceptions of the person, of society, of reasonableness and of justice as fairness as a whole would eventually be accepted as fixed points, precisely because they are obvious to every careful observer of western political culture and thought. But looking at the history of philosophy and the authors cited in this chapter, we can see that this has not happened yet and expecting that it will happen seems hopeful at best.

Maybe it is true that there are some widely shared essentials (values, ends, basic views of good social order) embedded in the political culture of western constitutional democracies. So we might assume that there is a kind of social truth that corresponds to the building material the constructivist is seeking. But, apart from being difficult to substantiate empirically, this assumption does not get us very far. For even with regard to the choice and nature of the basic building blocks, such as the correct conception of a person, opinions differ greatly. Accordingly, William Galston argues early on against Rawls:

> "The conceptual foundation of the basic structure – free and equal moral personality – is supposedly addressed to the citizens of our society. But Rawls's reconstruction of justice as fairness does *not* invoke – indeed, it flatly rejects – the conception of the person underlying our beliefs and practices. There is little evidence to support – and much to refute – Rawls's hope that his conception of personality will prove acceptable to us once its implications are fully grasped. Yet his "constructivist" metatheory leaves him no other grounds of persuasion or verification."
>
> (Galston 1982: 516)

For a more societal perspective George Klosko writes on a similar note:

> "Among the many issues over which adherents of different views will probably disagree are the precise characteristics of free and equal persons. [...] In the absence of strong evidence to the contrary, there is little reason to believe liberal citizens will agree more readily about these issues than about other aspects of their moral views."
>
> (Klosko 1997: 638)

Even if we further grant that this kind of controversy could be overcome, there would still be the third dimension of pluralism to consider. As Moehler's analysis of the different modeling of equality in Rawls and in Harsanyi shows, even if theorists hold the same view about the

importance of some aspect, it is an entirely different matter to agree on the best way of modeling it. As any social scientist would tell us, how to best model something does not simply follow deductively from some aspect of a phenomenon or process under consideration. The same aspect can be modeled in different ways and a model can always be changed so that some desired outcome – e.g. certain principles of good social order – follows from it.
In social science there are established ways of testing the correctness or at least the usefulness of different models. Constructivism in HCM is lacking such a test. That is, there are no established standards for deciding whether some model is correct, besides broad demands of coherence and reasonableness. This is what I mean when I say that constructivism in HCM lacks a genuine methodology.

My main objection to HCM, according to the normative reading, is that it does not offer any means for dealing with the pluralism we have on the level of normative social theory. Quite the contrary. As we have seen above, everything hinges on constructing correct hypothetical choice models. But besides broad demands for overall coherence and reasonableness, there are no standards of correct construction. Thus theorists employing HCM are invited to simply reproduce their preconvictions at various points of constructing. In this respect, HCM is like what Stefan Fisher has called a garbage machine: The kind of garbage it spits out depends on the kind of garbage the theorist used for its construction and there is no agreement in sight regarding the right kind of garbage. (Fischer 2018: §3.2) Even worse, constructivism in HCM does not offer any help, let alone a methodology, to get any closer to an agreement on the right kind of garbage.
Rawls should have clearly seen and addressed this issue. He is very explicit about the fact that reasonable pluralism extends all the way to philosophical doctrines. (Rawls 1993: 36-37) So why would he think that a procedure such as political constructivism could converge toward an agreement regarding the correct social point of view? Or, sticking more closely to Rawls' theory, why would he think that the outcome of a political philosophy can be "freestanding", which requires as its inputs the very things on which a freestanding conception should remain impartial: "philosophical, and moral doctrines"? (Rawls 1993: 144) If there is plurality in philosophical doctrines this implies – at least inductively – the existence of plurality in what philosophers consider to be the most reasonable social point of view. Accordingly Jeremy Waldron remarks:

> "Important though Rawls's conception has been, we all know that there is barely a hand full of academic political philosophers who accept the original position idea as Rawls expounds it or his view of the principles and guidelines that would be accepted therein."

(Waldron 1999: 153-154)

There are at least two passages where it does seem like Rawls sees this problem, but thinks of it as something that we simply have to live with.[33] But I disagree. As I attempt to show in the following chapters, there are promising alternative ways of professing abstract normative theorizing in spite of pluralism on all levels.

2.3.2 Two: Illegitimate Abstractions

Let us turn to the second line of critique I present in this chapter, following from the empirical reading of HCM. Actually this second line of critique consists in two independent critical arguments. The first one works up to the rather blunt claim that, empirically speaking, it is evident that none of the hypothetical choice models presented so far model the reasons we have. The more important second criticism defends the systematic point that HCM cannot possibly succeed in modeling the reasons we have, due to the preference structure real citizens have.

Hypothetical Choice Models Falsified

If we follow the empirical reading and understand HCM as essentially modeling the reasons we have, an obvious first point of critique is the lack of effort made by the mentioned theorists to employ empirical inquiry in constructing and testing their models. Focusing on construction, especially Rawls may rightfully be criticized for mainly relying on the popular activity of "armchair sociology"[34]. Gaus on the other hand does provide an empirically informed account of social morality as a core building block of construction. Be that as it may, any theorist constructing a model may point to Milton Friedman's claim that it does not really matter what goes into modeling or how realistic the model itself appears to be, all that matters is that it can do some explanatory and predictive work. This defense, however, highlights that HCM according to the empirical reading should at least feature some kind of test of whether the model has actually been successful in modeling the reasons we have. Gaus and Rawls acknowledge this as a requirement of full justification, which is meant to ensure that what is justified in abstraction in a hypothetical choice model also remains justified when presented to regular

[33] 1) "If sound, these remarks suggest that in philosophy questions at the most fundamental level are not usually settled by conclusive argument. What is obvious to some people and accepted as a basic idea is unintelligible to others. The way to resolve the matter is to consider after due reflection which view, when fully worked through, offers the most coherent and convincing account. On this, of course, judgments may differ." (Rawls 1993: 53) 2) "As to how we find the correct procedure, the constructivist says: by reflection, using our powers of reason. But since we are using our reason to describe itself and reason is not transparent to itself, we can misdescribe our reason as we can anything else. The struggle for reflective equilibrium continues indefinitely, in this case as in all others."(Rawls 1993: 96-97)

[34] I owe this expression to John Elster (1992: 146).

citizens. (Rawls 1993: 285-287; G. Gaus 2011: 365-366) Unfortunately, they never make the effort to devise an actual test that could establish whether their models come close to meeting the requirement of full justification. But perhaps this kind of test is unnecessary because we already have a pre-test amongst normative theorists that pretty much settles the matter. This is because, as indicated by the above quotes from Brian Skyrms, William Galston, George Klosko, Wilfried Hinsch and Jeremy Waldron, present hypothetical choice models already get rejected on the level of normative social theory.[35] And since normative theorists are citizens who are also experts on these matters, their rejection should be sufficient to show that these models do not truly model the reasons we have.

This leaves us with the same conclusion as drawn from the first line of critique. Namely that pluralism on the level of normative social theory seems to disqualify HCM as a viable approach – although in both cases, the criticism presented does not rule out the possibly of some unknown hypothetical choice model eventually succeeding in achieving extensive approval amongst normative theorists. If this day ever comes, according to the empirical reading there would be a rationale for devising a more inclusive, society-wide test of the model and the resulting principles in question.

Only Theorists Left Alive

Irrespective of whether this day will ever come, in the following I put forward a more systematic criticism of HCM according to the empirical reading. This criticism defends the stronger claim that HCM cannot successfully model the reasons we have as citizens. In short, the problem is that HCM cannot do both: construct a conclusive hypothetical choice problem and accurately model the kind of preferences we have. In trying to do so nonetheless, hypothetical choice models make use of an illegitimate abstraction, abstracting from the very preferences they are supposed to be modeling.

In more detail, first recall that one important goal in HCM is the construction of a choice problem that is conclusive, so that substantial normative principles can be derived in theory. In order to be conclusive in this way, the social point of view modeled in HCM must identify generally shared preferences. Whereby "preferences" is meant to emphasize that it is not enough for the model to identify any reasons citizens might have. Rather, what is needed are shared, "overriding" reasons that show us what citizens generally prefer. (G. Gaus 2011: 335-336) Only such preferences can ground normative principles of justified social order in theory.

Note, second, that this seems to be an impossible task in light of the assumption of reason-

[35] In Rawls' terminology: Justice as fairness did not establish general and wide reflective equilibrium among normative theorists.

able pluralism and all the different social circumstances individuals find themselves in, so we should expect preferences to also be diverse. Intuitively, we would expect that some actual society either obviously shares certain preferences for some way of doing things, or such shared preferences simply do not exist. In both cases, modeling shared preferences would be pointless.
Third, consider the strategy for resolving this problem in HCM. It consists in an argument from abstraction that establishes overriding preferences by constraining all possible perspective to one single "normalized" perspective. (G. Gaus 2017) A normalized perspective and the preferences derived from it then render the choice problem at hand conclusive. For a simple example consider rational choice theory. Rational choice theory yields many conclusive predictions of individual behavior in given scenario. This is because rational choice theory specifies a very narrow perspective on choice problems which roughly states: *Do not talk to people but only ask yourself what option maximizes your own expected utility and do it!*
In HCM normalizing is achieved in different ways. Rawls allegedly achieves this by restricting the choice problem to a choice about primary goods and by having agents choose behind the veil of ignorance. These modeling devices constrain the choice problem and the agents facing it in a way that we approach a singular, conclusive perspective in the original position from which – at least according to Rawls – his two principles of justice are favored over other principles.
Gaus, in order to derive his rights agency and jurisdictional rights, does not employ modeling devices such as the veil of ignorance. Instead he uses a strategy one might call "the human rights argument" in order to construct a normalized perspective. This strategy consists in firstly, asking us to abstract from diversity, secondly identifying things we all share and thirdly, deriving universal normative claims from what is shared. In this case, Gaus is asking us to abstract from all the things the Members of the Public disagree about. Then he points out that all members of the public see themselves as agents. Finally he concludes that all members of the public must endorse certain principles safeguarding their agency.
Note, fifth, that in constraining perspective to one conclusive perspective, both theorist abstract from the very things they are supposed to model according to the empirical reading: the (pluralistic) reasons we have. This is so because the normalized perspective introduces constrains that only hold within the respective hypothetical choice model so constructed. In Rawls' theory this is rather obvious due to the peculiar nature of the original position. In contrast you might think that Gaus abstractions are less problematic, because we do in fact all care about our agency. This is perhaps correct, but not much follows from it. To see this consider that there might be many reasonable accounts of what the perspective of agency requires. Also this does not mean that there are not other important perspectives (freedom, equality, justice, what have you) – what if they collide? What are the trade-off rates? Do all people have the same trade-off rates? Does one and the same person have the same trade-off rates

over different social contexts? Essentially what I am trying to point out here is that it is not enough to argue that most or all people care about their agency. It would only be interesting if it was clear what agency is and that it is always of overriding importance. Perhaps Gaus has shown what agency means for his Memebers of the Public and that for them these thing are of primary importance. But we are not Members of the Public in the deliberative model. We are just people with divers perspectives.
Conclude at last, that according to the empirical reading, normalizing amounts to an illegitimate abstraction, because it abstracts away from the very thing it is supposed to be modeling: the (pluralistic) reasons we have. But "whatever else we wish away in our elaboration of ideal models [...] we should not wish away the fact that we find ourselves living and acting alongside those with whom we do not share a view about justice, rights or political morality." (Waldron 1999: 105) Otherwise the normalized preferences are very unlikely to pass the full justification test for successful arguments from abstraction:

> "[I]t must be the case that the deliberative conclusions are not overturned as the process of abstraction is undone and Members of the Public are again understood to be guided by their full set of evaluative standards. [...] In the end, to publicly justify must be to justify in terms of all the relevant evaluative standards. We wish to structure common moral life on terms that everyone – considering all that she holds to be important and relevant – has sufficient reason to endorse."
>
> (G. Gaus 2011: 336)

Thus, on the empirical reading I believe Rawls and Gaus have both failed their own test of full justification. More concisely, they have failed to model the reasons we have because this is precisely what their models illegitimately abstract from in order to get a conclusive result. As we will see below, Rawls has more or less admitted this by conceding that justice as fairness is one among several reasonable conceptions of justice. Gaus has also done so indirectly by criticizing normalization in public reason theory and neglecting agency rights after OPR.[36]

Returning to and summarizing my main critical conclusion here, HCM, according to the empirical reading, involves an illegitimate abstraction that results from the normalization of perspectives that is necessary for rendering the choice problem conclusive in theory. Therefore it is likely that any instance of HCM will fail the full justification requirement of successful arguments from abstraction.[37] That is to say that if the process of abstraction is undone, if, as

[36] Citations are provided at the end of the upcoming *Subsection 2.3.3.*

[37] "This requirement is immensely important: Unless the conclusion of the argument from abstraction can be affirmed in light of a rational and reflective free and equal moral person's full set of evaluative criteria, the abstract justification will be defeated by these other elements of his or her evaluative set." (G. Gaus 2011: 336)

it were, the veil of ignorance is lifted and the reasoning of agents is replaced by the reasoning of citizens, the conclusions of the model are unlikely to hold.
If my argument is correct, the theorist employing HCM only has two options: One, abandoning her model and starting to look for a different way of modeling the reasons we have. Two, abandoning the empirical reading and arguing that the unique theoretical perspective and the unique preferences of the hypothetical agents she has constructed are correct – irrespective of the diverse preferences actual citizens might have. Then, of course, we would be back to the normative reading and its problems.

In conclusion of both lines of criticism presented in this chapter, what we see is that HCM does not escape pluralism and thus cannot provide a conclusive point of view in theory. On the normative reading, this is due to an ignorance of persistent pluralism on the level of normative social theory. On the empirical reading this is due to the persistent pluralism on the level of citizens and their preferences that can only be avoided by means of an illegitimate abstraction. Therefore, in contrast to Michael Moehler I do not see the differences in HCM as a "promising starting point for future research". To my mind HCM does not contain but rather stands in the way of establishing a helpful methodology in normative social theory. Therefore we have good reasons to start looking for an alternative.

2.3.3 Diverse Theories and Changing Views

So far, I have mostly focused on John Rawls and Gerald Gaus as main proponents of contemporary social contract theory and HCM. And although this restriction is useful in that it helps us to focus and limit the discussion, it also omits the diversity and change in contemporary theory. In order to further exemplify the diversity and how the lines of criticism developed in this chapter apply differently to different instances of HCM, let me introduce one more pertinent theory: David Gauthier's *Morals by Agreement*.

Rational Choice Style HCM

Gauthier's hypothetical choice model is "The Initial Bargaining Situation".[38] As the name suggests, he uses bargaining theory to model different individuals, i.e. maximizers of expected utility, with different preferences struggling to find a rational agreement on common norms or principles. The core feature of the initial bargaining position is that individuals are situated

[38] Actually Gauthier develops not one but two complementary hypothetical choice situations: "The Initial Bargaining Situation" and the "Archimedian point", whereby the latter denotes the idea of constructing an impartial, "moral" choice situation, similar to the Rawlsian project of constructing an original position. Ultimately Gauthier's claim is that both perspectives lead us to choosing the same principles.

within a "moral free zone": a perfectly competitive market in which initial personal and property rights are assigned and all interactions are mutually beneficial.
The core normative claim Gauthier draws from this setting is "the principle of minimax relative concession". Simply and informally speaking, the reasoning behind choosing this principle is that bargainers – according to Gauthier – compare their concessions associated with each bargain to the best outcome they could possibly achieve, still within the framework of mutually beneficial agreement. Given this measure of their concessions, his claim is that bargainers with greater concessions will be likely to demand a better deal, whereas those with smaller concessions are more ready to grant it. Thus relative concessions would minimize and, in most cases, concessions would equalize.
What are the main differences between the initial bargaining situation and the other instances of HCM discussed in this chapter? Firstly, what sets Gauthier's project apart is his aim for a purely rational reconstruct of our normative practices. This is primarily reflected by his use of bargaining theory and modeling the choosing agents as maximizers of expected utility. Now, of course, by assuming a perfectly competitive market and mutual benefit, Gauthier is effectively introducing a kind of minimal fairness into his model. This baseline prevents something like a society made up of slaves and slaveholders from being the point of origin or outcome of bargaining. In spite of this minimal normative safeguard, Gauthier's assumptions are less expensive than Rawls' image of agents that choose principles for allocating primary goods behind the veil of ignorance, while being unable to attach probabilities to outcomes.
Secondly, a major difference in comparison to Rawls is that Gauthier's normative claims are more abstract and in that sense less substantial than Rawls' principles of justice. Gauthier's principle of minimax relative concession is a meta principle, guiding bargaining and choice in the initial bargaining situation. With his maximin principle Rawls does something very similar, but in his theory this meta principle of choice is only an intermediate step preceding his principles of justice that apply directly to (constitutional) social order. Gauthier does not take this step.

There is a pattern here: Theorists that stick closer to standard rational choice theory tend to use less demanding idealizations and abstractions, while also ending up with less substantial principles.[39] In doing so, they are very much subject to my first, but less to my second line of criticism. That is, all of these theories exemplify the irreducible pluralism on the level of normative social theory, whereas rational choice style theories are less likely to make illegitimate abstractions in order to get a decisive result. Which in turn leads them to less substantial normative principles.

[39]Recall for instance John Harsanyi's argument for the principle of utility maximization as presented in *Subsection 2.2.2*, or consider Michael Moehler's argument for a "stabilized Nash bargaining solution". (Moehler 2018)

Further, theorists emphasizing the rational choice perspective seem to be more reluctant to construct a viewpoint that collapses the plurality of preferences into one single preference order. This is not surprising because in rational choice theory it is commonplace to take people and their preferences as they are. Accordingly, besides demanding consistent preferences, argumentation within individual utility functions is not permissible. As a result, the principles defended by these theories, such as solutions to bargaining problems, are highly abstract and a long way from actual norms of social order.

The general conclusion here is that there are of course different strands in how to construct hypothetical choice models. This nicely illustrates pluralism on the level of normative social theory. It also shows that my second line of criticism does not apply equally to all instances of HCM.

Habermas' Hypothetical Constructivism

Jürgen Habermas' constructivism is another interesting oddball from the perspective take in this chapter. I endorse many aspects of his theory of deliberative democracy and especially his critique of Rawls:

> "Philosophy shoulders different theoretical burdens when, as on Rawls's conception, it claims to elaborate the idea of a just society, while the citizens then use this idea as a platform from which to judge existing arrangements and policies. By contrast, I propose that philosophy limit itself to the clarification of the moral point of view and the procedure of democratic legitimation, to the analysis of the conditions of rational discourses and negotiations. In this more modest role, philosophy need not proceed in a constructive, but only in a reconstructive fashion. It leaves substantial questions that must be answered here and now to the more or less enlightened engagement of participants, which does not mean that philosophers may not also participate in the public debate, though in the role of intellectuals, not of experts."
>
> (Habermas 1995: 131)

In the preceding chapter I will also argue that as theorist we should take a more modest role, engaging only in reconstructions from existing social normativity and normative thought, while leaving most substantial questions to be answered here and now to the deliberations of citizens. Nevertheless, I do depart from Habermas' conception of how the "reconstruction" is to be undertaken. The reasons for this departure point us to the methodological core of Habermas' contributions to normative social theory. Fortunately, here we can rely on Markus Patberg's take on presenting Habermas constructivism in a concise manner:

> "[A] rational reconstruction aims at revealing the rational core of social practices. Furthermore, the reconstructive method rests on the assumption that an intuitive 'knowledge' on behalf of the participants as to the rational core of their shared practice is constitutive for the respective social action context. The implicit thesis is that certain practices would have to break down without the existence of the idealizing assumptions. The reconstruction targets those presuppositions, then, that are inevitable for the preservation of a practice. From this stems the explanatory function that rational reconstructions may fulfill. The idea of a constitutive function of counterfactual assumptions can be illustrated by means of the example quoted above, of citizens that cast their vote, motivated by their ideal picture of democracy, although political scientists explain to them how the electoral system renders their vote irrelevant. The inevitable presuppositions may be reconstructed by way of simulating a rational discourse as to the meaning of the practice in question."
>
> (Patberg 2014: 511)

> "[...] in this way, on the one hand he [Habermas] explains the persistence of these forms of social interaction and on the other hand the reconstruction [...] leads to the explication of the normative substance of the analysed practices and ultimately results in the formulation of a standard which may be applied with critical intent to the correlating practice."
>
> (Patberg 2014: 512)

There is a lot to digest here. An obvious and important difference to several of the other approaches discussed in this chapter is that Habermas' constructivism starts out with a social practice found in social reality – e.g. voting. This, I will argue in *Section 3.1*, is the entry point to a promising alternative to HCM that I call *Embedded Constructivism*. Simultaneously, Habermas' approach depicts a kind of *hypothetical yet empirical* constructivism. It is hypothetical because the normative substance of a given social practice is reconstructed by means of an imaginary procedure of rational discourse. Hence, although the reconstruction is about some actual practice and its participants, it is done by Habermas, sitting in his armchair.[40] It is empirical in that the hypothetical procedure is meant to reveal implicit, counterfactual presuppositions that actual people have. These idealizations allegedly explain why people uphold the practice in question (e.g. voting) *and* also yield the normative standards against which the practice in question may be judged.

[40] Hypothetically, every theorist has an armchair to theorize from.

I fully endorse Habermas' focus on particular social practices. I also believe that the meaning of such practices usually includes normative standards that may be unveiled or "reconstructed". Nevertheless, I fear that a hypothetical reconstruction of these standards runs into the same problems as hypothetical reconstruction in HCM according to the empirical reading. Specifically, I do not see why we should be confident that the hypothetical procedure of an imagined rational discourse should necessarily unveil the reasons that any actual citizens have for participating in some practice. Neither do I see why the actual reasons people have for upholding a practice are necessarily relevant for the standards that should be used for evaluating the practice. Finally, I do not think that the idea of hypothetical discourse leads to any determinate outcomes, but rather creates ample space for the theorist to stipulate whatever outcome her intuitions suggest.

Consider the example of voting. People may engage in this practice for all kinds of reasons (e.g. habit/internalization, social pressure, enjoyment, belief in certain benefits of electoral democracy, belief in the intrinsic value of voting, etc.). How could some imagined discourse explicate the actual reasons that motivate the persistence of this practice?

Further, assuming that people are actually motivated to participate in a popular vote by some counterfactual idealization, say, they believe that voting ensures freedom of the press, and that hypothetical discourse would reveal this (although I do not see how), why should this counterfactual predisposition, this illusion upholding the practice of voting be the basis for any normative standard of voting? In this scenario people were clearly mistaken for believing that voting per se has anything to do with freedom of the press. Of course, in a rational discourse, their mistake would be corrected, but this simply goes to show that the outcomes of hypothetical discourses and actual motivations, reasons or dispositions are two different things.

Finally, what about the idea of a rational discourse about the meaning of voting itself? Does such a discourse have a determinant outcome? Is voting primarily about expressing political equality, or about establishing a fair procedure, or about facilitating accountability? These are difficult questions that perhaps escape any definitive answer.

Overall, I believe Habermas' reconstructions are not about the actual reasons as to why some practice persists or what people value about this practice. If this was really the core concern, we would be better off using a method such as deliberative polling. (James S. Fishkin 1991) What is really going on is that Habermas is simply telling us what he thinks are the best reasons in favor of the practice in question. He is reconstructing a rational and normative standard for a given practice of interest as *he* sees it manifest in concrete actions, norm texts and institutions. (Patberg 2014: 511) This is all well and good, but it is not the sophisticated method he is selling it for. Thus, we should not be surprised if other theorists reconstruct

different standards or if actual citizens do not care for them.

Changing Views

Besides these differences in theorizing, I do see a general tendency in the literature to abandon the quest for constructing universal principles of good social order in theory. In respect to Rawls, Fred D'Agostino notes:

> "Indeed, what seems to have happened is that Rawls abandoned, during the course of his long career, both the goal he set for political theory and the fundamental modeling device that he adopted as a basis for pursuing that goal. In particular, Rawls in effect abandoned the idea that political theory ought to and could successfully aim at the identification of a public conception of justice fit to order competing social claims."
>
> (D'Agostino 2018: 30)

Perhaps speaking of abandonment is too strong, because Rawls defends justice as fairness, including the original position and the two principles of justice, as the most reasonable conception throughout all of his works. But he clearly conceded that there is a family of reasonable conceptions and "[o]f these, justice as fairness, whatever its merits, is but one."(Rawls 1997: 774)

In Gaus' work after OPR something similar has happened. The idea of a deliberative model is still present but has lost importance. In his 2016 *The Tyranny of the Ideal* it does not appear at all. Rather, what takes center stage is the worry that constructions of a single decisive perspective ("normalizing") in respect to the problem of choosing principles of good social order are misguided in a society of diverse perspectives. (G. Gaus 2016: IV.1.1, 2017) Accordingly, Gaus ignores his earlier argument for agency rights as human rights and restates his case for the importance of jurisdictional rights, not as a consequence of the deliberative model, but as a practical necessity of a diverse society in need of stable social order. (G. Gaus 2016: IV.2.4) So although I may still be at odds with Gaus regarding the question of whether the deliberative model was a case of illegitimate normalization in the first place[41], this disagreement remains of less importance, as Gaus' position becomes more practical and inconclusive in respect to substantial claims of good social order.

Overall, I do see a spreading renunciation of HCM.[42] Taken together with the criticism presented in this chapter, I hope the attentive reader is now sufficiently motivated to start elaborating an alternative.

[41] Gaus seems to uphold the opinion that it does not. (G. Gaus 2016: 23)

[42] See also Wilfried Hinsch (2018).

2.4 Concluding Remarks *Chapter 2*

At this point we can streamline and summarize the critical argument developed in this chapter and look at it concisely. Essentially my argument has been this:

1) HCM is presented by its proponents as a remedy to the problem of reasonable pluralism: Not only is an agreement between all citizens difficult to picture as an actual event, it also seems impossible in principle due to the fact that citizens hold different but equally reasonable opinions on matters of good social order.

2) In order to solve the problem of reasonable pluralism, proponents of HCM construct hypothetical choice models, i.e. theoretic social points of view. This construction builds on basic ideas of a cooperative society, a person, reasonableness and so forth, that are explicated and arranged into one conclusive choice problem.

3) Methodologically speaking, HCM is neither a case of thought experimenting, for its logic is deductive rather than experimental, nor a case of scientific as-if modeling, for it does not model any actual social process. It is rather a kind of constructivism that, depending on whether we follow a normative or empirical reading, rather models how we should or how we do think normatively about social order.

4) According to the normative reading, everything in HCM hinges on the *correct* construction of the hypothetical model. At the same time we observe extensive and lasting pluralism on the level of normative social theory, consisting in fundamental differences on all levels of construction. Since HCM does not include a method for dealing with this pluralism, it invites the construction of ever more models, reproducing theorists' differences in (philosophical) predispositions and their disagreements.

5) According to the empirical reading, HCM essentially *re*constructs the reasons we have in a way that leads to a conclusive outcome in the shape of norms or principles that are preferred by all. In order to do this, however, HCM makes use of an illegitimate argument from abstraction by constructing a single perspective that abstracts from the very things that are meant to be modeled: The diverse perspectives and reasons we have.

6) In conclusion, HCM either employs an illegitimate abstraction from what it is supposed to be modeling, or it is committed to providing the one correct or most reasonable construction, while reproducing the pluralism we have on the level of normative social theory. We therefore have good reasons to focus on an alternative way to proceed.

Limitations and Clarifications

Right away I would like to stress again the limits of this argument. I have not claimed that the difficulty in establishing the correctness of models of hypothetical choice implies that HCM cannot possibly succeed. In spite of pluralism on the level of normative social theory, we may still achieve convergence or even agreement on the correctness of one or several models of hypothetical choice. In other words, it might turn out that the pluralism in normative social theory is not as reasonable or as deep after all. But at least inductively, considering the lessons of past and present theorizing, this scenario seems highly unlikely.

Further I would like to stress that my criticism is not aimed at the pluralism in philosophical views held by different theorists itself. Deep reasonable pluralism is the hallmark of academic reasoning that has not truly escaped philosophy and formed a separate discipline. Therefore I do not criticize pluralism in normative social theory, but that HCM offers no way of making any progress in face of these divisions.

I also wish to prevent further misinterpretation of my argument. One might be led to believe that my argument was meant to establish that *one* correct hypothetical choice model will probably never be agreed upon. But I would be missing the point that there are really many and potentially many correct models of hypothetical choice aimed at settling different fundamental questions of social order. From this perspective, different theorists do not present directly competing models of hypothetical choice, but rather different models for solving different problems. Note, however, that this is no objection to my argument. Because for every single instance of HCM, both lines of criticism provided in this chapter still apply.

All in all, I am convinced that the hypothetical turn in social contract theory was a mistake. Theorizing about the reasons we would have under unique conditions, specified in theory, does not enlighten us about the reasons we do have.

Chapter 3

Embedded Constructivism

In this chapter I begin to spell out the positive side of my argument. I do this by introducing three ideas, which respond to the problems of HCM discussed in the previous chapter and also serve as building blocks for the kind of theory I will eventually propose in the next chapter. Generally I ask the reader for some good will and patience while considering these building blocks. This is because, while the ideas I put forward should of course be intelligible and recognizable as sensible reactions to the shortcomings of HCM, they are not fully explicated and integrated into a coherent whole. This is why I refer to them as "building blocks" that still need to be integrated into one theory in *Chapter 4*.

The overall goal of the chapter is twofold. On the one hand, the goal is to formulate an alternative approach of how to best profess normative social theory in response to failings of HCM. On the other hand, we need a range of suitable ideas – "building blocks" – that exemplify how applying this alternative approach might play out when trying to come up with an alternative theory of publicly justified social order.

Note that an alternative is always an alternative to something but not to everything. Consequently, what I am proposing in this chapter is not necessarily an alternative to all the approaches and theories so far employed in normative social theory. Neither do I have the resources to provide a discussion of all theories that might offer an alternative to HCM in one respect or another. One has to start somewhere. I take it that my critique of HCM, i.e. a critique of a core part of the most influential works in the field (namely those of John Rawls), is a legitimate starting point.

3.1 A Descriptive Start

One dimension of the pluralism on the level of normative social theory described in the last chapter are the different conceptions of the basic problem of justification. More precisely,

theorists confront us with different things that they consider to be the object of justification. Rawls speaks of justice, Gauthier of morality and Gaus of social morality. These notions are not entirely distinct, nor are they identical. This points to a more general and a more specific problem. The general problem is the awkward, yet persistent lack of a unified understanding of central concepts in normative social theory. What terms like 'justice', 'morality' and 'the political' mean and how they relate to each other, is settled coherently at best within specific theories. Between different theories, however, such terms remain pseudo technical terms – i.e. terms that are used frequently, as if they were important, well-defined basic concepts of some discipline, whereas they really just keep on fueling endless debates.[1]
The more specific problem with respect to theories of justified social order is that the object of justification is usually itself a theoretic, normative conception – typically a conception of justice or morality.

In Rawls' theory, the starting point and object of justification is a normative conception of justice: justice as fairness. Further, the core building material of his hypothetical choice model are notions of rationality, a person and society. To my mind, this is a problematic starting point, because such conceptions are also only specified in theory and are potentially the source of endless debate. In this section I argue that a descriptive, or ideally, empirical account of social order as the object of justification offers a better starting point for normative theorizing. The advantage consists in having a fixed point we can hold onto and look back to when professing normative theory. This fixed point is not "fixed" in the sense that it is beyond doubt and debate. However, since the matter of actual social order is an empirical one, it produces debates of a different domain. My hope is that due to the established standards and methods in the empirical domain, it will turn out easier to handle the matter of pluralism on the level of normative social theory. That is, I hope it will turn out easier to agree on an appropriate descriptive or empirical, rather than a theoretic and normative conception of the proper object of justification. If this is so, starting out with a descriptive or empirical conception of social order as the object of justification would relieve us of much disagreement and controversy. It could further induce more convergence in normative theorizing as a whole. Because if theorists start out with some descriptive or empirical conception of social order, their views on an appropriate framework for thinking about it normatively are more likely to converge.
In this first section I proceed by explaining in more detail the problem involved in starting out

[1]Debates that, taken together, do make one wonder whether a concept such as justice or morality "is itself merely a WEIRD invention: a historically recent, culturally parochial, psychologically uninteresting honorific used by different communities to commend whatever their favored subset of normativity happened to be, and by different researchers for whatever purposes were rhetorically convenient." (Kelly n.d.)

with normative, theoretic conceptions of justice and morality. Secondly I introduce my favorite candidate for an empirical account of social order: Cristina Bicchieri's account of social norms. Thirdly, I extend her account to also cover legal and moral norms and, fourthly point out the overall reasons for starting with an account of social norms.

3.1.1 A Bad Start: Conceptions of Justice and Morality

The first step of this section consists in explicating the problem with conception of morality or justice found in HCM and how we can avoid it. Now, of course 'morality' and 'justice' can be used in a sensible manner and I don't wish to argue that they are inherently bad or confused. I rather believe that they are closely intertwined with a misguided intellectual practice of assuming that there must be a mysterious sphere of social normativity, where some truth about the right kind of norms can be discovered. It is this line of misguided thinking that has provided for much of the motivation to engage in HCM in the first place.

Gaus' Conception of Social Morality

To exemplify this point, we return to Gerald Gaus' notion of social morality in his *The Order of Public Reason*. I choose this example because, from my perspective of arguing for an empirically well-founded conception of social order, Gerald Gaus' notion of social morality is quite progressive. At the same time, Gaus eventually digresses into the construction of a hypothetical choice model - his "deliberative model" - which is precisely what I wish to avoid. Thus my intention here is to carve out where I believe he went wrong.

In his section on *Moral Rules as Social Rules*, Gaus maintains that moral rules must be actual social rules. (G. Gaus 2011: III.10) That is, they must be rules that can and do govern actual behavior. As such rules, they depend on the beliefs and expectations people have in a given group. Most importantly, they depend on whether individuals believe that a rule is actually being followed by others and whether others also expect them to follow it. Social rules also depend on psychological mechanisms such as scripts, which allow people to categorize a given situation and activate the respective appropriate behavior. For instance, tipping the waiter after having been served food in a restaurant.

Overall, Gaus' conception of social morality pays close attention to accounts of how real world norms work. He also endorses the implication that from such a perspective we should expect there to be many different workable social rules and thus different moralities to be found. Thus, Gaus' conception of social morality already tends to a major problem in normative social theory, namely that

> "[t]oo many moral philosophers and commentators on moral philosophy [...] have

> been content to invent their psychology or anthropology from scratch and do their history on the strength of selective reading of texts rather than more comprehensive research into contexts."
>
> (Darwall, Gibbard, and Railton 1992: 188-189)

The perspective of social rules is one of two sides to Gaus' conception of moral rules. Not surprisingly, the other side is a normative perspective of justified rules. What is surprising, however, is the terminology Gaus employs in presenting the normative perspective on rules, especially given the descriptive perspective of social rules that I have just summarized. Gaus introduces the normative perspective on rules with the notion of "true morality" and a "moral point of view":

> "What Baier calls a 'true' morality is one that passes certain tests of impartiality and common acceptability. Testing existing social moralities on the basis of such considerations is to evaluate them from 'the moral point of view'."[2]
>
> (G. Gaus 2011: 177)

Now, I do understand the urge of philosophers to introduce a critical perspective on the rules we have. After all, this is precisely what I am doing in this inquiry. What I do find surprising, however, is an account of actual and diverse social norms to be followed by the notion of a special set of "true" norms. Of course, Gaus only speaks of "true" morality here in reference to the work of Kurt Baier. Nevertheless, to my mind a more cautious notion – such as "justified", "optimal" or the like – would have been helpful. Further, why would we call a normative perspective on social morality a "*moral* point of view"? From the point of view of social "moralit*ies*" we would expect that there are as many moral points of view as there are moralities or moral beings. So, if morality denotes a diverse set of social rules, and if one were to seek a critical perspective on such rules, why would one also denote this as a "moral" perspective? It only makes sense if one were to assume that two sets of rules existed: social moralities and metamorality that allow us to evaluate moralities.
This seems to be precisely what Gaus assumes, because further below we learn that true morality is to be found in another mysterious entity called "transcendent morality":

> "Morality, we may think, must provide a perspective that transcends the social order so that we make claims regardless of that order. [...] [A]lthough the core tasks that morality performs require that it be embedded in a social order, we must

[2]Gaus' reference here is to Kurt Baier (1958): *The Moral Point of View: A Rational Basis of Ethics.*

> be able to stand back from our social institutions and take the perspective of what, we might say, "morality itself tells us." The moral principles that transcend the social order are, however, highly abstract and subject to wide-ranging interpretive controversy. Witness the idea of human rights: prior to attempts at codification in the international order, they functioned primarily as transcendent moral claims which, while having some content, are subject to endless controversy about what they are, and what they require of whom."
>
> (G. Gaus 2011: 180)

There are several things going on here. So let me try to dissect the goals, assumptions and approaches pointed to in the quoted passage. Firstly, Gaus suggests that the goal of a theoretic and normative inquiry into the nature of morality is to carve out general, "transcending" claims regardless of specific instances of social order. What kind of claims these might be remains unspecified in the quoted passage. However, the example of human rights as well as the fact that Gaus eventually moves on to construct a hypothetical choice model, producing certain abstract rights, suggest the answer: The claims he is talking about are what I called substantial claims of good social order in the preceding chapter. These are claims about the content or structure that actual social orders ought to embody.

Secondly, since the "we" in Gaus' text usually refers to him and the passive reader, it is Gaus himself, in his capacity as a normative social theorist, who is to produce these claims. So the underlying assumption is that substantial claims of good social order can be established in theory. This, to me, is quite a questionable assumption to make, especially in a world where we find a plurality of instances of social order ("moralit*ies*"), lively discussion about their faults and virtues as well as individuals holding different respective preferences. In face of this colorful picture, it should at least be considered as an open question whether universal claims of good social order can be found at all and whether this can be achieved in theory.

Thirdly, Gaus further simply assumes that carving out substantial claims of good social order in theory must be possible, because besides the visible moralities we have, another transcended morality must exist – somewhere. This is a strange assumption, because the ontological status of what is assumed remains a mystery. And again, to me there seems to be nothing tangible in the world that would render this assumption plausible.

Admittedly, the mystery somewhat dissolves when we consider Gaus' and similar projects in their entirety. Then the common and less mysterious answer seems to be that transcended morality exists in our shared nature as rational and reasonable social beings. Accordingly, Bernard and Joshua Gert (2020) hold that all normative accounts of morality "[...] refer to a code of conduct that, given specified conditions, would be put forward by all rational people."

And this seems to be precisely one of the guiding ideas behind the theories of Harsanyi, Rawls, Gauthier and Gaus, as presented in the last chapter. I am not trying to argue here that this project is doomed to fail. Perhaps our common human nature as social beings does provide a strong rationale for endorsing certain norms. But initially, the question should remain open as to whether this is the case, what the right methodology (or methodologies) would be to achieve this, and whether anything substantial could be determined about the content of such norms from a theoretical perspective.

The problem with Gaus and Rawls I am pointing to here is that they fail to see that these matters should be considered open questions. By introducing their conceptions of morality and justice, they are assuming a separate normative sphere to the matter of justified social order, besides the actual people, preference, discussions and instances of social order. In this sphere, something like transcendent or universal insights into good social order are possible and this sphere can be mentally reached by a theoretic construction of the appropriate perspective of shared human rationality or reason. Thus, they are simply assuming right from the start that their approach, involving the construction of hypothetical choice model, must be possible.

Rawls and the Priority of Justice

In Gaus this questionable assumption is introduced with the move from "social morality" to "transcendent morality". With respect to Rawls this point is best exemplified by an analysis of his famous opening statement in *A Theory of Justice*:

> "Justice is the first virtue of social institutions, as truth is of systems of thought. A theory however elegant and economical must be rejected or revised if it is untrue; likewise laws and institutions no matter how efficient and well-arranged must be reformed or abolished if they are unjust."
>
> (Rawls 1971: 1)

This statement is indeed a powerful opener. Rawls implies that the question of justice leads to a, if not *the*, central normative perspective on social order. But if we allow ourselves a moment of reflection, Rawls' opening statement turns out to be of mere rhetorical substance. Note first that having just started out reading *A Theory of Justice*, it is difficult to assess the statement at all. It primarily hinges on what 'justice' means and it is the core objective of any theory of justice to eventually tell us what it means. Second, note that the statement is meant to display an obvious truth, while this is obviously not the case. There could be other things – such as stability, security, provision of subsidence level goods and rights – which people might consider more important than justice, whatever it is (even more so, if the core value of

justice is fairness, as Rawls later tells us).[3] Third, note that the question remains open as to whether there is a unified thing called "justice", which one could develop a theory of. We know that there are things we call "social institutions". And it is further a familiar and plausible next step to take a normative perspective on such institutions and ask about the right kind of institutions. However, the notion that there is a whole other normative dimension to this – namely the dimension of just institutions, principles of justice and choosing such principles – is anything but obvious. To exemplify this point, consider the very real possibility that 'fairness' – Rawls' core value of justice – denotes nothing more than a set of different local norms regarding appropriate distributions in different contexts. (Bicchieri 2006: 83) Hence 'fairness' may simply be the name of a category for socially approved, distributional principles. Something similar could said about 'justice'.

None of this is to say that Rawls' rhetorical trick at the beginning of *A Theory of Justice* renders his theory as a whole invalid. But it does show that Rawls simply assumes right from the start that above and beyond social institutions, constituting a society's basic order (e.g. constitutions, constitutional courts, basic law or norms visible in political culture), there is a normative meta level and we can access this level in virtue of theorizing and reasoning. And although his readers may ultimately be convinced that this is plausible, it sure seems a strange assumption to start out with. For all we know, 'justice' is a complicated concept with different meanings. Generally speaking and looking at the history of this concept, 'justice' is probably best understood as a general label for discussions on the question of what we owe to each other in society. (Schramme 2006: 23-24) That this concept denotes or is helpful in conceptualizing any more than a broad realm of thought is something to be shown by Rawls' theory, not something he can take for granted right from the start.[4]

There are of course motivations behind the quest for justice and transcendent morality that we, as social beings, can all relate to. As such beings, born and socialized into an existing social order, we know the experience of strong rules that seem to be of great importance and categorical nature. We also might think that there must be some general, reasonable principles of good social order and it sure would be nice if someone could tell us what they are. To be this someone who can discover and offer principles of good social order is in itself an appealing

[3] "Justice is but one of many virtues of political and social institutions, for an institution may be antiquated, inefficient, degrading, or any number of other things without being unjust. The notion of justice is not to be confused with an all-inclusive vision of a good society; it is only one part of any such conception." (Rawls 1963: 73)

[4] Another example of Rawls simply assuming that what he is trying to do must be possible is his fourth "general fact of political sociology and human psychology" in the restatement of his original theory: "We add, then, a fourth general fact: that the political culture of a democratic society that has worked reasonably well over a considerable period of time normally contains, at least implicitly, certain fundamental ideas from which it is possible to work up a political conception of justice suitable for a constitutional regime." (Rawls 2001: 34-35)

position of importance and power. But as normative theorists we should not let this lure us into building the belief in universal normative principles right into the conceptions of our basic objects of inquiry. Because if we do, we will assume the existence of a mysterious and and perhaps nonexistent normative sphere.

A Descriptive Account as a Remedy

My goal up to this point was to show how conceptions of morality and justice, used by Gaus and Rawls, lead to the construction of hypothetical choice models. In more detail, these conceptions of morality and justice have a build-in assumption of a separate normative sphere, which can be accessed by means of the proper theoretic point of view and ultimately yields substantial claims of good social order. Thus, these conceptions lay the groundwork for the construction of hypothetical choice models and thereby for the problems pointed to in the last chapter. Essentially, 'morality' and 'justice' are highly contested theoretical concepts with a blurry relation to reality, and as such give rise to hypothetical choice models of the same nature.

The remedy, I suggest, consists in putting notions such as morality and justice aside for a change and replacing them with a descriptive account of social order as the object of justification. In reference to my discussion of HCM and hypothetical constructivism in *Chapter 2*, this is to say that I suggest to change the building material that goes into the construction. The advantage of doing so is that it allows us to carefully differentiate between a descriptive account of actual social order and the normative perspective of ideal order. An example for a descriptive account of social order would be to specifically consider the justification of constitutions. But also a more general conception such as Gaus' conception of social morality or Rawls' basic structure[5] of society can serve this function. Insofar I do think that both theorists start out in the right direction by offering general descriptive accounts of social order. The wrong turn they take, then, is the blending of their descriptive accounts of social order with ideals about the right kind of rules a society should have. As explicated above, in Gaus this happens with the move from "social morality" to "transcendent morality". In Rawls this happens when he informs us that his theory is restricted to the analysis of the basic order within a "well-ordered society", already governed by a conception of justice. (Rawls 1971: 4-5; 8-9) In both cases, the descriptive conception of social order is mixed with an ideal notion of social order. This blurs the object of justification in that we are no longer certain where and whether said entity exists. It also invites endless debates about the correct normative conception and how to theorize about it.

[5] "By the basic structure I mean a society's main political, social, and economic institutions, and how they fit together into one unified system of social cooperation from one generation to the next." (Rawls 1993: 11)

Therefore, I suggest starting out with a purely descriptive conception of social order, which is meant to capture something that actually exits in ordinary societies and that does not contain any normative connotations about ideal order. A good first indication for whether a descriptive notion of social order has been formulated is to see whether it is neutral with regard to normative qualifications of its content: That is, a descriptive conception does not allow one to qualify any rule or institution as "just" or "unjust", "moral" or "immoral", "justified" or "unjustified". Simply put, a descriptive conception of social order does not contain any notion of good social order. A second important characteristic of a descriptive conception is that it does not presuppose whether any substantial normative perspective on it can be constructed at all.

A descriptive conception is also accessible to empirical investigation – at least in principle. Ideally, the descriptive conception is not only accessible to empirical methods in principles but is already fully operationalized. Being empirical in this sense has several advantages. One of them is that an empirical conception is very precise in meaning because it defines some general, possibly abstract concept in terms of some set of specific observations. Further, my hope is that having such a conception will reduce controversies over the correct conception. This hope in turn is based on the impression that, with an empirical conception, it is easier to decide where we have substantial disagreement and where we are simply concerned with different instances of social order. And where substantial disagreement remains, it is usually also easier to attain more agreement on the most appropriate conception because we can evaluate different conceptions in virtue of their fit with the evidence collected from the phenomenon in question. Last but not least, starting out with a descriptive and empirical conception increases our chances of ending up with a normative theory that is practically meaningful and relevant. That is, it increases our chances of coming up with a normative theory that offers orientation for societies and their orders as they exist today, because it never exchanges them for their idealized representations as the object of theorizing. Thus we are likely to avoid the problem – typically faced by theories employing HCM – of having to show why the outcomes of ideal theorizing matter at all to actual citizens.[6]

3.1.2 Biccheri's Account of Social Norms

There are of course many descriptive conceptions of social order that can serve as the starting point for a theory of normative theorizing. In particular, there are different descriptive accounts of different instances of social order. However, since I am concerned with the justification

[6]This problem of stability is the core motivation behind Rawls move to *Political Liberalism.* (Rawls 1993: XVii-XViii) See also Gaus' discussion on *The Stability of Abstract Rights Under Full Justification.* (G. Gaus 2011: 359-368)

of social order in general, I am in need of a general descriptive account. To this end, my favorite candidate is Cristina Bicchieri's theory of social norms.[7] Therefore, in what follows I summarize Bicchieri's conception and try to show that it is even more extensive than Bicchieri herself believes.

Social norms, according to Cristina Bicchieri, exist in the form of symmetric, conditional preferences and beliefs. In detail, the idea is that i has the preference to conform to rule R, if

(1) i believes that sufficiently many other individuals will also conform to R and if

(2) i believes that sufficiently many other individuals normatively expect i to conform to R.[8]

(3) For some individuals it must also be true that i believes that others may sanction a failure to conform to R.

Now, condition (1) is a straightforward existence condition of norms. But conditions (2) and (3) really set social norms apart from other norms, such as conventions and other behavioral patterns Bicchieri calls "descriptive norms"[9]. This is because the introduction of normative expectations and sanctions brings us into the world of social normativity where normative obligations are felt and expressed. One important feature of social norms is that they are - once established - behavioral equilibria. This means that if (1), (2) and (3) are the case for a set of individuals, no individual norm follower has an incentive to violate the norm. Rather, the beliefs of all individuals will be self-fulfilling and the norm will thus be stable. A second important feature of social norms is that they are not necessarily good or efficient. Quite to the contrary, there are lots of bad norms around. Bicchieri's example of choice is a stable social norm that prescribes violent behavior among gang members because all gang members believe that the other gang members will commit acts of violence and expect this of other gang members, while every individual gang member detests violence. (Bicchieri 2006: 180) A third key point on Bicchieri's account of social norms is that it amounts to an *as-if* explanation, common in economics and game theoretic modeling. Such explanations do not necessarily

[7]Interestingly, Gaus also considers Bicchieri's account and offers a definition of social morality that is consistent with it. (G. Gaus 2011: 170; 181-182)

[8]In her formal definition of social norms, Bicchieri calls this condition the "normative expectations" condition (Bicchieri 2006: 11). She further specifies that these expectations could be understood in a normative *and* in an empirical way (Bicchieri 2006: 15). I find it difficult to see why merely empirical expectations of others about what I will do should be important to me. Perhaps in situations where I have a preference for coordination.

[9]Descriptive norms are all kinds of behavioral patterns that emerge because some people like to conform to standards of behavior they observe. This is for instance the case with fashionable clothing. People might wear certain clothes because they believe them to be fashionable. Conventions on the other hand exist when individuals have a symmetric preference for coordinating on some outcome. A standard example is the problem of deciding upon which side of the road to drive on. (Bicchieri 2006: 29-42)

describe what actually goes on in people's heads. So it does for instance not assume that people always (consciously) deliberate through conditions (1) to (3) before they act. Nevertheless her as-if model of social norms is an explanatory and predictive model amenable to testing.[10] Bicchieri further goes far beyond providing an as-if explanation by connecting her rather formal model with a psychological account of norm following. Here Bicchieri argues that norms are embedded in a complex web of social knowledge. Such knowledge provides individuals with categories and scripts which offer an interpretation of a given situation, including roles and appropriate behaviors.

> "For example, once I cast the person I am facing into the category 'waiter', a script about what happens in restaurants is primed, followed by the prediction that this person will come to my table with a menu, take orders, bring food, and so on. A script may also contain rules and expectations about the restaurant client's behavior, including ways of addressing waiters and tipping policies."
>
> (Bicchieri 2006: 81)

Bicchieri goes on to show that her model of social norms is coherent with a range of empirical findings in psychology and behavioral economics. The upshot of these discussions is that her account of social norms has the advantage of being general and flexible at the same time. It is flexible in that it does not simply claim that in situation S people always follow social norm N. It rather explains how norm following is one important element in individual choice besides others. Thus the existence of a norm does not necessarily mean that it is (always) followed. Further, what is prescribed can vary greatly between different cultures and social contexts. This flexibility in Bicchieri's account of social norms also allows for its robust generality. By generality I mean that social norms can help to explain very different kinds of behavioral patterns, ranging from classic pro-social behavior such as trust, reciprocity and fairness, to cases that do not have such positive connotations, such as the violent behavior of gang members.

In conclusion, Bicchieri provides us with a rich account of social norms. Social norms are artifacts of human interactions. They exist embedded in shared interpretations of social situations - e.g. a costume party - and shared beliefs about the regular and expected behavior in this situation - e.g. wearing a costume. If these interpretations and beliefs are widespread and symmetric, a norm has a good chance of being a self-enforcing behavioral equilibrium manifested in actual behavioral patterns - e.g. everybody shows up at a costume party wearing a costume and feeling that they are behaving in the correct manner.

[10] "I am not claiming here that mine is a realistic model of how we reason, but [...] I maintain it is a fairly good explanatory and predictive model, because my definitions are operational and their consequences are testable." (Bicchieri 2006: 48)

3.1.3 Incorporating Morality and the Law

So far so good. But why should we focus on social norms within an inquiry into the justification of social order? At this point in particular one might be wondering whether social norms are too narrow of a conception in this context. Bicchieri herself fuels this worry by distinguishing between social, legal and moral norms. On this point, however, I am more with Briam Skyrms:

> "I think of the social contract not as some monolithic unitary pact, but as an assemblage of norms. Norms are conventions that are backed by sanctions. Sometimes the sanctions are codified in the law and enforced by government. Sometimes norms are not explicit, but rather implicit in practice, and sanctions take the form of some type of social pressure."
>
> (Skyrms 2016: 1087)

Looking at things more closely, we can see that legal as well as moral norms essentially work through the same mechanism of conditional preferences as social norms, or so I argue. Thus it might still be sensible to distinguish moral, legal and potentially other norms, but only as different instances of social norms.
In the following I firstly rejoin the laws and social norms and secondly morality and social norms from a functional perspective.

The Leviathan Reconsidered

Let us begin with the relation of legal and social norms. By legal rules I mean codified laws, produced and backed by state institutions. Bicchieri says right at the beginning of her book that she does not consider norms that are explicitly designed and enforced by political institutions:

> "I call social norms the grammar of society because, like a collection of linguistic rules that are implicit in a language and define it, social norms are implicit in the operations of a society and make it what it is. Like a grammar, a system of norms specifies what is acceptable and what is not in a social group. And analogously to a grammar, a system of norms is not the product of human design and planning."
>
> (Bicchieri 2006: ix)

This analogy between grammar and informal norms is puzzling, because the grammars we use today are codified scripts which are discussed, intentionally designed and taught through formal institutions such as schools. Sure, if I were to only learn a language according to some

textbook, I would miss many linguistic practices which go beyond and even contradict textbook grammar. But I would still learn an awful lot about a language by looking at a grammar book. The same is true for societies and their laws: The written laws do not tell us everything, but they sure are (ever more) important in understanding the normative structure of a given society. So right from the start, I do not see why we should ignore written laws if we wish to understand a society's "grammar", or why we should attempt to draw a sharp line between norms that are of intentional design and those that are not.

Metaphors aside, I think the more fundamental problem is how Biccheri and others view the state. According to this view, laws are different from other informal social rules because laws are enacted and enforced by a mysterious animal, abundant of authority and power.[11] I cannot attempt a comprehensive historical reconstruction of this idea. One obvious point of origin is Thomas Hobbes' *Leviathan*. Recall the original frontispieces of Hobbes' famous 1651 book, depicting a gigantic monarch who, on closer look, is made up of a great number of individuals, all looking up at the head of their sovereign king. For Hobbes it was crucial that the sovereign is a powerful and politically unrestricted authority, because on his account, only such an authority can sustain social order. To my mind, the key features of this picture are, one, the fact that the Leviathan is physically made up of individuals, which reflects Hobbes' revolutionary individualistic perspective. It follows that the power of the Leviathan is obviously nothing other than the combined powers of all of its subjects. Two, it is curious that the individuals do not look at each other, but rather at the head of their ruler. This in turn suggests that the power of the sovereign can only be exercised if the subjects pay attention to her commands.

Unfortunately, it seems that many thinkers have also taken the perspective of the subject looking up to her ruler and thus fallen prey to the illusion that state authority is something we can take as a given. They really should have taken the perspective of the outside observer who is puzzled by the sight of so many individuals being effectively coordinated by the commands of one. From this outside perspective, provided already by the frontispiece of *Leviathan*, it appears obvious that state authority, viewed as an independent authority, enacting and enforcing laws, is an illusion. All we can see from the outside is a set of individuals who are following, for some strange reason, the commands of one ruler, rendering the one ruler powerful. Arguing from this perspective that people follow the commands of the ruler because the ruler is powerful clearly results in a circular argument. So what then explains political authority?

[11] One seminal example is John Austin's command theory of law, according to which laws work as commands (i.e. enforced demands) of a sovereign government, habitually obeyed by its citizens. However, *why* citizens habitually obey some government remains a mystery. (Austin 1832: esp. I)

From the Leviathan to the Republic of Beliefs

Inspired by the phenomenon of corruption and state failure, Kaushik Basu (2015) offers an elaborate argument of why state authority in the form of effective laws is also just a behavioral equilibrium, such as a social norm.[12] Basu's argument can be summarized in three steps.

The first step consists in recognizing that state enforcement – at least up to now – is not some automated machine that just hands down punishment and rewards. Enforcement is carried out by people (police officers, clerks, judges, prison guards and so on) no different than the followers or breakers of laws.

The second step is to acknowledge that if laws function through the coordinated behavior of a group of individuals (citizens + state functionaries), then laws also can only guide actual behavior if they are an equilibrium, sustained by coherent beliefs about what others will do and what they ought to do. To see this, Basu suggests looking at states that are successful and states that fail to guide their citizens' behavior effectively through laws. On the face of it, enacting a law is just writing words on a piece of paper. In some countries this action systematically changes how people behave. In failed or corrupt states on the other hand, enacting a law does not have this effect. At an abstract level, what makes the difference between the two cases is people's beliefs. If people believe that what the government signs into law is a good indicator for what the others - citizens *and* functionaries! - will expect and do the next day, this will provide them with powerful reasons to respect the law. If, on the other hand, many people believe that laws are commonly ignored by citizens and enforcers alike, the enactment of any law will not make the desired difference in observed behavior.

The third step of the argument consists in the conclusion that if laws can only be effective as equilibria between all the individuals involved, then what the government does by enacting an effective law is to make one feasible equilibria *focal*. You can think of this as someone putting a spotlight on one of several options and thereby coordinating the actions of all the observers. Again, this will only work if many people believe that others will actually treat the spotlighting as a good indicator for what will generally be done and expected.[13] Hence, powerful governments are not powerful primarily because they have a lot of police officers, guns or tanks, but because the commands and laws they utter are believed to be good indicators of what people (citizens and functionaries) will do the next day.[14]

[12]Ken Binmore (2005) criticizes in a similar vein how many thinkers have assumed the existence of an external enforcement agency, and argues that we can think of effective social order only as comprehensive equilibria.(Binmore 2005: 148-149)

[13]The underlying idea here is to model government, effectively governing through law, as solving a coordination problem with several possible equilibrium solutions. This idea is already developed in a similar way by Robert Cooter (1998). Basu's and Cooter's works are in turn based on Thomas Schelling (1960) analysis of "focal points" in coordination problems.

[14]It is similar to the curious phenomenon that some teachers are treated with respect and others are not by the

The relation between social norms and the law has been a topic of great interest for quite some time. Thereby a popular theme seems to be that there is an important difference between formal and informal social rules – i.e. between laws and social norms. (Posner 1997) More recently, however, the insight that laws are just a way of selecting between different possible behavioral equilibria is gaining popularity. (Sunstein 1996; Carbonara 2017) Basu's argumentation is a case in point, showing that, on a fundamental level, legal norms work in the same way as social norms. Namely through a system of symmetric, coherent expectations or beliefs that self-enforce a certain behavioral pattern. Therefore I consider legal norms to be a subcategory of social norms.
Their specific characteristic is that they are deliberately produced, enacted and enforced through an institutionalized social system, whereas other social norms are intentionally or unintentionally produced and enforced by an informal social system. Further, laws rely on a shared meta-norm that explains why the subjects look up at the Leviathan and translate words on paper into action and behavioral patterns. That such a meta-norm is necessary is already hinted at by Max Weber's account of descriptive legitimacy (especially the notion of "Legitimitätsglauben"). (Weber 1964: 158) In greater coherence with the terminology used in this inquiry, we may follow Garry Mackie (2018) and say that effective laws, i.e. laws that produce intended behavioral patterns, require a social norm of legal obedience.

The Illusion of Strong Rules

According to Bicchieri's definition, moral norms are norms that are followed because individuals have strong personal normative beliefs – such as "Killing is just wrong!" – which motivate them to act accordingly. These normative beliefs could be conscious, propositional mental states, but they might as well become habitual and internalized. Then, "[t]hese beliefs become an independent motivation to conform, as deviations are often accompanied by guilt." (Bicchieri 2017: 32) Bicchieri also suggests that norms that are supported by personal normative beliefs are themselves unconditional – i.e. they do not depend on expectations about what others do or what others think one ought to do:

> "[S]ome may argue that there really is no difference between social and moral norms, others would object. My objective here is not to examine the nature of morality. All I want to call attention to is that there is an element of (social) unconditionality

same group of students. From the students' perspective, obeying or not obeying the teacher are both possible equilibria and if one of the two is in place, they are usually quite stable. One important skill of teachers is thus the ability to establish the equilibrium in which they are the respected leader. And it is a tragedy if a teacher fails to do so, because once it is the established belief that some teacher is generally ignored, it is extremely difficult to revert the dynamic of self-enforcing beliefs.

> to what we take to be moral rules that is not present in social norms, in the sense that one's personal moral convictions are the primary motivator of one's actions, and such convictions overwhelm any social considerations."
>
> (Bicchieri 2017: 31)

There is a tension here in Bicchieri's theory. On the one hand, she is claiming that moral norms are distinct from social norms because they are affirmed by people's personal normative beliefs and thus hold unconditionally. On the other hand, her own findings are quite clear that empirical expectations about what others generally do are decisive for the existence of a norm and generally trump beliefs about what should be done. (Bicchieri and Xiao 2009) She also acknowledges that moral norms are not unconditional in every scenario. If, for instance, (social) conditions change radically, people may no longer follow moral norms, although they still hold the respective personal normative beliefs. (Bicchieri 2017: 32)

This is an important observation. Our actual societies can and have fallen prey to widespread nasty practices, undermining all the high-held principles some call morals. Sure, people may have strong (internalized) normative beliefs that motivate following (or rejecting) some norm. But this does not show that there are any actual norms that hold unconditionally. Another problem here that we will discuss in more detail in *Subsection 4.1.3*, is that we do not know how internalization of norms actually works. It seems probable that humans have a dedicated system – a "norms system" – for picking up and internalizing norms we observe in our social environment. It is further likely that internalization is highly responsive to the common behavior we observe, but it is unclear how it responds to reasoning.

The main lesson here is that 'morality' is an elusive and problematic concept. We often use it to refer to the phenomenology of norms; especially to the phenomenon that certain norms feel as if they demand unconditional compliance or are supported by weighty or universally valid reasons. This phenomenology should, however, not deceive us into believing that norms some call "moral norms" are truly unconditional and work differently than social norms.[15] Further, we should acknowledge – as Bicchieri does – that there is no particular domain of morality. Essentially, "[...] what makes something a social or a moral norm is our attitude toward it." (Bicchieri 2006: 21) This means that anything could be a moral norm. There is not "[...] any empirically important – let alone well-delineated – phenomenon deserving of being partitioned off as morality. No subcategory of norms makes up a psychologically distinctive or cooperatively indispensable set of moral ones." (Davis and Kelly 2018: 19)[16] Chad Van Schoelandt (2018)

[15] Generally, we should keep in mind that the best explanation for a given phenomenon does not need to present itself in the same narrative or be based on the same concepts and ontology as its explanandum. Phenomena labeled "moral" exemplify this point. (R. Kelly 2017: 352-353)

[16] This quotation belongs to an illuminating discussion in a recent issue of *Behavioral and Brain Sciences*. The

adds that much of the practice of "morality", specifically having appropriate reactive attitudes of resentment towards another, makes sense precisely if we assume a web of widely shared social norms to be in place.

Overall, I believe the phenomenon "that there is an element of (social) unconditionality to what we take to be moral rules" is just that; a phenomenon experienced from the first-person narrative of a norm-follower.[17] It does not mean that the norms between us truly hold unconditionally. If I had to explain why we are often under the illusion of unconditional norms, I would firstly point to the fact that there are fairly universal *types* of norms that exist in almost all human communities.[18] Thus there are fairly universal, important reasons to have such norms. Secondly, I would conjecture that the illusion of unconditionality is an artifact of cultural evolution. Specifically I presume that it provided a certain evolutionary advantage to have stable norms (especially the type of norms just mentioned) hammered into us by religion or codes of honor. But in politically advanced societies, i.e. orders of public reason with strong institutions, we do not need these illusions. Further, 'morality' continues to obscure rather than aid attempts of understanding how norms actually work in human communities.

In summary, my aim in this discussion of the relation between social, moral and legal norms was to show that legal and moral norms function like social norms through a web of symmetric, coherent beliefs or expectations. This is not to say that we cannot distinguish different social norms. Thus, one could define moral norms as those social norms that demand unconditional compliance and are supported by the strongest emotions or most universal reasons. And one could think of legal norms as those social norms that are intentionally produced and enforced by an institutionalized system, which in turn depends on a meta-norm of effective government.

3.1.4 Why Social Norms?

So far I have argued for leaving normative, theoretic conceptions of social order aside and instead starting out with a descriptive conception as the object of normative theorizing. Further I have introduced Bicchieri's account of social norms as a viable candidate for a general theory of justified social order, including moral, legal and other norms. I now turn to a more systematic elaboration of the reasons in favor of basing a theory of justification on an account of social

discussion revolves around a proposal from Kyle Stanford (2018b), who presents objectivity or externalization as a key phenomenon of morality. Having been criticized by several commentators for presenting morality as a separate normative sphere, Stanford (2018a) eventually admits that this was a mistake and restates his proposal in terms of "norms" instead of "morality".

[17] The quoted fragment stems from the passage in Bicchieri (2017: 31) quoted above.

[18] Most human societies have norms that prohibit killing, physical assault and incest, as well as norms of fairness and assistance. But how theses types of norms work out in detail is quite diverse. (Sripada and Stich 2006: 281-282)

norms.

The first reason is that social norms are the kind of things that can trigger the need for justification. This is because unlike descriptive norms and conventions, social norms are supported by normative expectations, which require individuals to do or not to do something. They thus create an obligation of compliance, and when faced with such a social obligation one might rightly ask: What is the justification for this obligation? Social norms exist where people are not only guided by their personal interests, but also by social pressure, obligation and the threat of punishment.[19] This is where the problem of justification comes in, because it consists in the nontrivial task of rejoining individual reasoning with the social obligation to follow a norm.

A second and closely related reason for relying on an account of social norms is that it allows us to systematically rejoin the empirical perspective of explaining actual social behavior and the normative perspective of justifying norms and obligations by means of reasoning:

> "I believe different people may have different reasons for compliance that extend beyond the standard reasons given by many social scientists, namely, that we fear punishment when we disobey a norm. [...] I would argue that another reason for compliance is the desire to please others by doing something others expect and prefer one to do. [...] A third reason for compliance with a norm is that one accepts others' normative expectations as well founded. In this case, sanctions have no weight. If I recognize your expectations as reasonable, I have a reason to fulfill them. I may still be tempted to do something else contrary to your expectations, but then I would have to justify (if only to myself) my choice by offering alternative good reasons and show how they trump your reasons. This need to offer a justification (to myself as well as others) signals that I recognize others' expectations as cogent. [...] Fear and the desire to please are powerful motives, but they imply that a norm would only be followed in circumstances in which either there is monitoring of one's actions and sanctioning is possible (as in repeated interaction) or there is some way to ensure that one's action is acknowledged by the people one wants to please or else has a noticeable effect on their well-being."
>
> (Bicchieri 2006: 23-24)

So there are different reasons for the effectiveness of social norms. Taken together with the demand for justified social order this, quite naturally, suggests a normative perspective on social

[19] Bicchieri adds a note of caution in respect to distinguishing norms: "The neat boundaries I drew between descriptive norms, conventions, and social norms are quite blurred in real life: Often what is a convention to some is a social norm to others, and what starts as a descriptive norm may in time become a stable social norm." (Bicchieri 2006: 38-39)

norms: Namely, that we refer to those norms that are followed only due to fear of punishment or the desire to please others as "bad" norms. So bad norms are unjustifiable norms that create extra costs of enforcement and the experience of authoritarian rule. Whereas we call "good" norms those that are supported by the reasons we have. Even in absence of surveillance, individuals will follow good norms because they have their own good reason to do so. Hence, "good" (or "justified") norms provide us with the kind of social order we want: made up of norms that are justifiable to us as citizens that do not require excessive enforcement.

This is not meant to imply that there are only justifiable and unjustifiable norms. In practice the reasons for or against existing norms are often not considered explicitly. That is to say, individuals may simply follow some inherited, internalized norm without ever having given much thought to it. You could call norms that are followed in this way habitual norms. They may turn out to be justifiable or unjustifiable norms on due reflection, but as it stands, we simply do not know whether one of the two categories would be fitting.

Bicchieri's account further points to the realistic expectation that in social reality we will most likely not find a neat distinction between justified and unjustified social norms. Rather we should expect to find norms that are followed, or partly followed, for a mixture of different motivations. Thinking of a scenario where people in society only follow norms because they have their own good reasons to do so is an ideal state and such an ideal eventually has to come to terms with the realities it is aimed at. Doing so is one core aim of the approach I am developing in this and the following chapter.

A third reason for the reliance on an account of social norms is that it is empirically well-founded. As already pointed to above, my hope is that starting out theorizing with such an account will reduce debate and divergence in normative theorizing. Now, of course Bicchieri's account is not the only proposal for understanding norms.[20] Thus, also on the descriptive and empirical side there are and continue to be debates about the best conception of norms. However, in this domain we have a clear reference point of testing the helpfulness of different conceptions: actual behavior. The helpfulness of descriptive accounts can in principle be determined in respect to their ability to explain and predict actual behavior. Further, if the account in question is not only descriptive but empirical (operationalized, predictive), we can turn to established empirical methods and start testing. Bicchieri's account has the advantage that it is already empirical in this way. It is also virtuous in that it provides us with a formal model and a rich explanation that extends from the individual, psychological to the societal level. Crucially, it also forms a central node of the emerging theory of social order, which I

[20] One fairly recent alternative that focuses on the normative attitudes we have toward social rules is proposed by Geoffrey Brennan et al. (2013). The authors claim to have constructed a counterexample ("The Chastians") that shows how Bicchieri's example leads to counter-intuitive outcomes. However, I agree with Kai Spiekerman (2015) that Bicchieri's account handles this scenario just fine.

will lay out at the beginning of *Chapter 4*.

This concludes the first section. Overall, my suggestion here is to leave theoretic conceptions of morality and justice aside for a change and try to build a theory of justified social order around a descriptive and empirical account of norms. On the one hand, this suggestion is motivated by pointing to the problems that come with normative conceptions of justice or morality as the object of normative theorizing. As seen above, theoretic and normative conceptions of justice or morality as employed by theorists such as John Rawls and Gerald Gaus assume the existence of a separate normative sphere. These conceptions are, one, highly debatable ideas with a blurred relation to empirical reality and, two, they have led those theorists to the intellectual practice I criticized in the preceding chapter: constructing hypothetical choice models.
On the other hand I have stressed the advantages that come with a descriptive and empirical conception of social order such as Cristina Biccheri's account of social norms. The advantages are, firstly, that it allows us to start with a tangible, empirical object of justification. Thus we can hope to have a clear grasp of the kind of thing(s) we are talking about and will not get into abstract normative debates right away. Secondly, I argued that Cristiana Bicchieri's account of social norms rejoins the empirical perspective of actual norms and the normative perspective of justified norms. Thirdly, I have argued that Bicchieri's account is particularly virtuous in that it combines a formal and testable model with a rich psychological background theory of norm following.

3.2 An Open-Ended Ideal

In the previous section I insisted that we should start out with a descriptive account of social order. This, however, may provoke the question of how then a normative perspective on social order can get off the ground and if it could lead us anywhere interesting. After all, and on this point I agree with theorists employing hypothetical choice modeling, a theory of justified social order should at least offer a critical perspective on the social orders we have. This is the minimum a normative theory has to achieve in this context. So how do we get to a critical perspective when starting out with a descriptive account of social order?
In the previous section I already hinted at the first step in giving an answer. It consists in distinguishing between justified and unjustified social norms and saying that the former are the good ones, because justified norms have certain desirable features pointed out in *Subsection 1.1.2*: they are likely to be stable, effective, pleasant and intrinsically desired. However, this intuitive way of identifying good social order with justified social norms does not give us much. In fact, it seems we are at a dead end, because for all we know so far, anything and nothing

could be the content of justified social order.
As we have seen in *Chapter 2*, theorists employing HCM avoid this dead end by constructing social points of view which allegedly identify certain shared and overwriting reasons for endorsing a comprehensive agreement on principles of good social order. With these principles, such theories are able to give at least an abstract answer to our second research question: *What is the content of justified social order for some given community?* In doing so, however, these theories also carry a heavy – and to my mind unbearable – burden of justifying the basic assumptions and their arrangement needed for constructing conclusive choice problems in theory.
But what if we could somehow avoid this burden altogether and use the original idea of a society having a justified social order without theorizing about its content? This is where the idea of an open-ended ideal comes in, because it allows us to do precisely that: Thinking about some group of individuals actually having achieved some generally desirable state – i.e. having established a justified social order – while leaving open the order that materializes when being in this ideal state. The most prominent instantiation of this idea is the notion of market equilibrium in western economic theory. I proceed by firstly discussing the example of market equilibrium in order to carve out its main features. Secondly, I attempt to transfer these insights into the realm of political theory.

3.2.1 Market Equilibrium as an Open-Ended Ideal

Political and economic theory are very similar. Both fields deal with an important domain of social life. Thus the subject matter of both fields is essentially institutions and social relations. Accordingly both fields are facing fundamental normative questions of social order: The question of *What are the right rules?* in the case of politics and the question of *Who gets what?* in the case of economics. Further, theorists in both fields are confronted with a problem of pluralism: They have to assume that individuals hold very different preferences regarding the appropriate political regulation and economic distribution. Also, in both fields theorists lack access to comprehensive knowledge of these preferences. Nevertheless, at least in western tradition, both fields have come up with their respective general answer to their basic normative question: democracy and market economy.
However, just as there are striking similarities, there is also a striking difference. Market theory has achieved a considerable degree of conformity, precision and coherence. The nature and dynamics of market economy are taught and discussed in the same – even formalized – terms all over the world, wherever people are susceptible to these kinds of ideas. Whereas democracy remains an elusive and much contested notion. Another way of looking at it would be to say that economic theory has established a stable, comprehensive paradigm, while political theory

has not. Why this difference exists is a fascinating question in itself, but unfortunately far beyond the scope of this inquiry. Here I want to pick out one central element of the market paradigm – the notion of market equilibrium – and harness its virtues for the task of suggesting a better way of professing normative political theory.

Some Basic Economic Theory

In order to get a better understanding of market equilibrium and its relation to the idea of an open-ended ideal, I will introduce some basic ideas of economic theory. The aim of this excursion is not a critical evaluation, but an explication of the approach to normative theorizing employed in this tradition. Let us begin with the typical framing of the basic problem at hand in economics:

> "Economics is the study of how societies use scarce resources to produce valuable goods and services and distribute them among different individuals. If we think about the definitions, we find two key ideas that run through all of economics: that goods are scarce and that society must use its resources efficiently. [...] Efficiency denotes the most effective use of a society's resources in satisfying people's wants and needs."
>
> (Samuelson and Nordhaus 2009: 4)

Simply put, economics is about the distribution of material things and services. Thereby economists are guided by one central ideal: (Pareto-) Efficiency in the satisfaction of individual economic wants and needs. That economists share this basic concern for efficiency is quite remarkable. In political theory it is hard to imagine that several theorists could actually settle on one guiding ideal in the realm of politics. This is of course not to say that having one guiding ideal, such as efficiency, is necessarily uncontroversial. But it sure is helpful to have such an ideal for establishing a paradigmatic theoretical framework.

Now, the procedural answer to the original distributional problem in western economics of course consists in the market mechanism:

> "A market economy is an elaborate mechanism for coordinating people, activities, and businesses through a system of prices and markets. It is a communication device for pooling the knowledge and actions of billions of diverse individuals. [...]. Yet in the midst of all this turmoil, markets are constantly solving the what, how, and for whom. As they balance all the forces operating on the economy, markets are finding a market equilibrium of supply and demand. "
>
> (Samuelson and Nordhaus 2009: 26-27)

The ingenuity of this answer consists in actually not answering the original question of *Who gets what?*. Economists rather suggest a mechanism – the market – and a general ideal state – efficient equilibrium – while not making any substantial claims on the correct distribution. Thus economics solves the problem of pluralism by proposing a system that processes the different preferences there actually are, whatever they are. Thus market theorists needs some knowledge of the general structure of preferences economic agents might have, but they do not need to know the preferences a particular person might have and neither do they have to think about what they would be in some idealized and abstract hypothetical choice model. All they need to think about is the open-ended ideal of a Pareto efficient state and what kind of social mechanism could produce it.[21]

Lessons From the Calculation Debate

At this point it is useful to make a short excursion into the history of economic theory. For it was not always the established paradigm of western economics that market equilibrium cannot and should not be determined by anyone or anything other than the market mechanism itself. Rather this position is the outcome of a long-lasting debate between those who believe in the necessity and possibility of an economy of central planning and those who do not. The origin and heydays of this debate date back to the nineteen-twenties and thirties and came to be known as the *socialist calculation debate*. Essentially the quarrel was between, on the one side, the emerging Austrian school of economics – represented by Ludwig von Mises and Friedrich Hayek – and, on the other side, a range of different proposals on how economic planning in the absence of private ownership of the means of production could be possible.[22] In a nutshell, Mises and Hayek argued that only markets, by means of the price mechanism, can adequately process all the relevant preferences and information in a complex, dynamic economy. Thus, every attempt of a central planner to externally calculate the correct equilibrium distribution is very likely to fail. This position remains at the core of western economics up to this day and explains why economists are so reluctant to give any substantial answer to the question of *Who gets what?*.

Interestingly, Hayek further pointed out that the idea of market equilibrium is neither something that is meant to be somehow calculated by experts in the first place, nor something that actually materializes at all at some given point in time in an ever-changing economy. Thus

[21] One could of course be more critical here in respect to basic economic theory. For instance, one may point to the substantial assumption about preferences in basic market theory that consumers always prefer more to fewer goods. Interestingly, Samuelson and Nordhaus (2009) state the example that at some point more ice cream just makes you sick, while not presenting us with a model (e.g. of negative utility) that could capture this possibility. (Samuelson and Nordhaus 2009: 85)

[22] An excellent contemporary overview is provided by Hayek's own (1935) anthology of the different contributions, including his own critical position.

the static image of equilibrium should be understood as a hypothetical idea and not as part of an explanation of how markets actually work. Actual market dynamics should rather be understood as a flowing river or an evolutionary process, not a static state of equilibrium.[23]

My impression is that normative political theory also needs more of a "calculation debate". That is, a debate about how and to what extent normative claims can be made in theory, given an ocean of unknown preferences. My criticism in the preceding chapter can be seen as a small contribution to this debate. One way of re-framing this criticism in light of what we have learned from Hayek and the calculation debate would be to say that theorists employing hypothetical choice modeling are engaging in illegitimate central planning: To some extent they try to calculate what all would agree to, although they lack the required data (individual preferences) to do so. And even if they did, there is probably no unique set of principles that everybody agrees on hidden in this date in the first place. Just as there is no unique equilibrium for any actual market to be reached.

These considerations also remain the backdrop of the ideas put forward in this and the following chapter. For now, let me summarize what we have learned about the nature of market equilibrium and explain why I see it as an "open-ended ideal". The *open-endedness* of market equilibrium understood as an open-ended ideal consists in three things. Firstly, it consists in the fact that we do not and cannot know the correct social order, in this case the correct allocation of goods and services, for any given society at any given point in time. Secondly, over time the unknown correct order is constantly changing, because circumstances (preferences, technology, availability of resources etc.) continue to change. Thirdly, the matter of establishing the correct social order is a matter of an actual, open-ended social process – in this case market exchange. Taken together, these three things mean that the ideal is open ended in that it leaves the problem of correct social order up to an open-ended social process. But what then is the substance of such an ideal? It is a description of general and desirable features some unknown and ideal state of social order should have. In the case of market equilibrium, the generally desirable feature is Pareto Efficiency in the satisfaction of consumer wants. Thereby, the idea is not to formulate an actual end point of market exchange. For, as Hayek has already pointed out, in the real world exchange will probably continue indefinitely and never reach perfect equilibrium. Market equilibrium is about envisioning the general features of an ideal outcome to the end of providing guidance for the design of the mechanism producing actual outcomes. In that market equilibrium, understood as an open-ended ideal, is rather procedure-oriented than outcome-oriented: It envisions an abstract and ideal outcome in order to guide the social mechanism or procedure producing the actual outcomes. In a

[23]See Karen Vaughn (2013) for an overall analysis on Hayek and his understanding of 'equilibrium'. See Hayek (1935: 226 ff.) and Hayek (1981) for some relevant passages from Hayek himself.

nutshell, an open-ended ideal provides a general and abstract understanding of a desirable social state, which points us to the kind of mechanisms for bringing about such a state in actual society.

There is something very appealing about the way market theory approaches the problem of good social order in face of pluralistic individual preferences. What I mean is that instead of thinking about what people want and what they should get, economics has focused on a social mechanism that can deal with incorporating, processing and balancing individual preferences. As a normative guide for shaping this mechanism they have relied on an ideal that is open-ended in respect to the matter of correct distributions. This is an elegant approach, for it shows how normative theorizing can be productive without speculating about shared individual preferences and the correct order justified by them. It rather allows the normative theorist to accept people and their different preferences as they are and might be in the future, while also allowing her to make a contribution in the form of providing an open-ended ideal.

3.2.2 Gaus on Justified Social Order as an Equilibrium

Since having an open-ended ideal has worked out well for economics, I suggest that we start thinking about a similar conception in the realm of normative political theorizing. In the case of economic theory, what is brought into equilibrium are individual preferences over different economic goods and services. In the case of normative political theory a somewhat analogous ideal would be an equilibrium of well-considered individual preferences over norms, so that stable norms are also justified norms, i.e. "stable for the right reasons". (Rawls 1993: 458 ff.) Right off the bat, we can see that an obvious difference between the two kinds of equilibria consists in the kind of underlying preferences. In the case of economics, typically relevant preferences are driven by individual expectations of consumption or profit. In the political realm however, we need to allow for a much broader conception of preferences that go into the equilibrium dynamics.[24] This is because in the case of political choice, people are expected to take into account all kinds of considerations. So they might be thinking about economic benefits, but they are probably also concerned with their visions of the good live and the good society when they form preferences over different possible social rules. A further related difference between the economic and the political realm is that in the former decisions are private, whereas in the latter they are public. More precisely, economic decisions are typically about what an individual wants to have or to invest, whereas political decisions are typically about norms that all have to live by.

[24]This is not to say that economics should not also generally work with a broader conception of preferences. Here I am simply acknowledging the fact that this is typically not the case in classic economic theory. For an alternative conception see for example Paul Ekins' (1992) account of a "progressive market".

These brief preliminary reflections imply that equilibria of justified norms are based on a broader notion of preferences and that the adjustment and balancing of these preferences has to take place in a different arena, better suited to public decision-making than the market. In order to further pursue the idea of norms as justified equilibria, let us return to Gaus' *The Order of Public Reason*, for it already covers a lot of ground to this end.

An Optimal Eligible Set of Justifiable Norms

In the last chapter I criticized the "deliberative model" elaborated by Gerald Gaus in OPR as an instance of an essentially unhelpful approach to the problem of justified social order. In Gaus' theory, the deliberative model is one of three main elements, constituting his theory of public reason in OPR. The other two elements are his account of social morality and his equilibrium solution for selecting moral rules. In the first section of this chapter I pointed to Gaus' account of social morality, arguing that it contains much important work in painting a more realistic understanding of the norms we live by, while also displaying the common mistake of assuming the existence of a questionable normative sphere at the beginning of normative theorizing. Gaus' equilibrium solution on the other hand is the one element of his theory that I fully endorse. In fact it captures quite well the idea of justified social order understood as an open-ended ideal. So let me offer a brief summary.

Gaus' argument in favor of the possibility of justified social order – or "justified moral rules" as he would put it – as a real-world equilibrium of reason rests on one core assumption: For the group of individuals in question, there must be an "optimal eligible set" of norms which are preferred by all to having no norm in place at all. That is to say that if some norm were part of this set, everybody had good reasons for endorsing it even if the norm in question were not the best norm by everybody's standards. This is because – qua being a member of the eligible set – we know that every individual would prefer this rule to not having any rule in place. Since all members of the optimal eligible set are also assumed to not be pareto dominated by any alternative, we also know that such a norm would be an equilibrium: Every individual would have his or her own good reasons for endorsing it and every intention of moving to a different equilibrium would be vetoed by at least one other individual in the group.

Gaus further provides an argument for why a group of individuals could converge on such a justified equilibrium. He calls it the "Kantian Coordination Game"[25]. Basically the idea is that people tend to coordinate on a rule due to an increasing returns dynamic. If for some reason a subgroup of individuals G coordinates on a specific norm X, any separate individual is drawn to also adopt X, even though he or she might prefer alternative rule Y, simply because adopting X allows him or her to cooperate with all the members of G. The larger G becomes,

[25]For a discussion of the actual game see Gaus (2011: 395-397).

the larger the incentive to adopt their norm X gets. If there is no equally popular contender, it is likely that the whole population will eventually converge on X.

> "This equilibrium is not only explanatory but justificatory; it does not simply explain why we arrive at a common morality, but that we have reached it justifies this morality (this rule) over other members of the optimal eligible set. How we have arrived at this rule is a combination of contingent history, moral ideas, happenstance, and the exercise of power. The route to it is path dependent. All these are important aspects of a social evolutionary account of justified morality, and all should be endorsed."
>
> (G. Gaus 2011: 418)

So the Kantian Coordination Game is a theoretical argument for the possibility of convergence on a justifiable norm, given the existence of an eligible set of potentially justified norms. But this argument remains silent on the question of why some society does in fact coordinate one X, and why it coordinates on X rather than on Y. Here Gaus holds that we have to look at the forces of real social dynamics such as path dependency, power and other historical contingencies. Here it is important to stress that he does not claim that moral rules always naturally evolve in this way. Rather, also distancing himself from Hayek, Gaus explicitly rejects the view that a rule is justified in virtue of having been selected as the positive rule through some social mechanism.[26]

A Testing Conception of Public Reason

So the evolutionary coordination on a norm as a behavioral equilibrium merely explains how such a coordination is possible. Whether the rule is justified depends on its passing of the test of public reason. This, on Gaus' account, essentially amounts to the question of it being a member of the optimal eligible set.[27] So far so good. But this leaves us with the vague and optimistic view that actual societies probably have a set of justifiable norms, either in the form of norms that are already established or that could be established in the future. As a consequence, we have no idea whether any actual norm is justified or not, rendering the whole account to be of mere academic interest. Therefore, Gaus provides us with a more

[26] "I have not postulated any mechanism such that orders that converge on a justified morality have some selection advantage over those that do not. We can remain agnostic about whether such a mechanism exists (it would be nice to think it did)." (G. Gaus 2011: 420)

[27] "The Deliberative Model explicates the moral point of view, and what is acceptable is any option in the optimal eligible set. That is the test. If x is in the optimal eligible set, then x as a current social rule is now the basis of a moral equilibrium: a rule that has been converged upon and can be freely followed, and whose authoritative nature can be acknowledged by each while consulting only her own evaluative standards." (G. Gaus 2011: 425)

substantial understanding of the appropriate test of public reason. Here the idea is that we can test existing or proposed social orders against the outcomes of Gaus' hypothetical choice model (the "deliberative model"). As a reminder, these outcomes are the idea of an optimal eligible set of justifiable norms and a set of basic abstract rights constricting the set. So any viable candidate for justified social order must embody some concrete version of these abstract rights.[28]

One example of how this test works out is Gaus' rejection of socialism. Here Gaus argues that, empirically speaking, systems allowing for widespread private ownership in the means of production are more effective in providing the kind of basic rights and freedoms we all want according to the deliberative model. Hence socialism will be dominated by individual preferences for systems with extensive private ownership and thus will not be in the optimal eligible set. (G. Gaus 2011: 511-521)

I do not wish to get into the details of this argument against socialism here, because I already reject Gaus' testing conceptions of public reason on a more general level. The reasons for doing so of course relate back to my criticism of hypothetical choice modeling presented in the last chapter. Let me capitalize on this opportunity to clarify and summarize my stance toward Gaus' account in OPR. On the one hand, I fully endorse his account insofar as it is consistent with the goal of developing an open-ended ideal. That is, I endorse the idea of an optimal eligible set of justifiable norms that probably exists for most given societies. In itself this idea is open-ended in that it does not contain any substantial claim on the right kind of norms. Further, it could also be understood as an ideal that is primarily procedure-oriented. Accordingly, an alternative version of Gaus' theory could focus on the practical question of what it would mean for an actual society to choose and establish a member of the optimal eligible set. But as I pointed out above, Gaus takes a different route. More precisely, in *The Order of Public Reason* he goes on to speculate about the principles – jurisdictional and agency rights – all "members of the public" would agree on in theory. I rejected this step in the last chapter because, one, we lack the resources in normative theory to establish the correctness of such claims and, two, such arguments from abstraction necessarily rely on illegitimate abstractions in order to render the choice problem conclusive in theory. As I put it above, deriving agency rights and jurisdictional rights form Gaus' deliberative model is an instance of illegitimate central planing.

In rejecting the substantial claims of Gaus' hypothetical choice model I also reject his testing conception of public reason, making use of these claims. I do however share Gaus' concern for

[28] "In all large-scale societies the conception of persons as self-directed agents is current and spreading; [...]. The Great Society is now a worldwide society, and its participants conceive of themselves as such agents. The morality of agency is thus today a universal, transcendent morality: all true moralities must accommodate the basic claims of agency. We may truly say that the claims of agency are human rights." (G. Gaus 2011: 430)

having some test in the first place. This is because, having an open-ended ideal of justified social order without such a test is rather unsatisfying – it leaves us with an ideal that does not have a systematic connection to the social order we actually have or could have. Thus it lacks practical meaning for us as citizens. One way to establish this link is to think about the mechanism that would tend toward producing justified equilibria in reality. This line of thought will continue to occupy us in the following sections.

3.2.3 Political Equilibrium as an Open-Ended Ideal

At this point we can summarize what we have learned about justified social order as an open-ended ideal from economic theory and Gerald Gaus' work on moral equilibrium and start modeling all of it into one coherent conception.
But first, let me remind us of our guiding justification principle of social norms (JPN) as presented in *Subsection 1.3.2*:

> **JPN**: A social norm N is justified to an individual i in society S governed by that norm to the extent that N being a positive norm in S is coherent with i's preferences, given that
>
> 1) i has formed well-considered preferences on social order.
> 2) N being a positive norm in S is strictly preferred by i to having no social norm governing the domain of N in S.
> 3) i is at liberty to openly reject the JPN in S.

The challenge in meeting the JPN stems from thinking about how it could be fulfilled by the norms of complex and diverse societies. Gaus' hopeful argument for the existence of an optimal eligible set claims that the challenge can be met. More specifically, he believes that there probably is a large enough benefit for all individuals to coordinate on cooperative norms and that the same benefit will create a bandwagon effect[29] for converging on a particular norm, which, once it is established, is then uniquely justified to everyone.
In order to have a real-world example of how this could work, think of the emergence and spreading of a messenger software such as WhatsApp. There are several such services we can use. Some people might really like WhatsApp, while others prefer alternative solutions. Nevertheless, everybody benefits from coordinating on one means of communication. If all your friends and colleagues were to use a different service, this would defeat the purpose of

[29]The notion that a new equilibrium norm is established through a bandwagon effect goes back at least to Cass Sunstein (1996).

having such a tool in the first place. For then you might as well call up everybody separately as you used to. So there is an obvious mutual benefit in coordinating on the same messenger and for some contingent reasons, WhatsApp has emerged as one of the dominating solutions. Given this fact of coordination we can assume that the dominance of WhatsApp is justifiable to its users, although it is neither objectively the best available messenger software, nor is it the preferred solution for all users. It is rather uniquely justified because it provides a desirable service (just as alternative platforms would) with the added advantage of being the established equilibrium most people are already using. Formally speaking, WhatsApp can be seen as a uniquely justified equilibrium to a coordination problem, whereby we are assuming that everybody benefits from coordinating on one solution and it is not the case that there is an alternative social network that all prefer.

Nevertheless, we are not in the world of social norms yet because so far using WhatsApp is just a convention with no normative expectations attached. But once this convention is firmly established, people's attitude toward it may change. This is because for someone already using WhatsApp, it is convenient that others do so as well, whereas it is rather annoying to have a friend or colleague who insists on being contacted via a different messenger or – god forbid – via phone call. Having to coordinate people over different technologies is simply more costly than doing so using one shared medium. Therefore, those who have coordinated on WhatsApp may, after having formed the empirical expectation that other people generally use WhatsApp, also form the normative expectation that others *should* use WhatsApp.

The important point now is that, as long as the original assumption of the coordination problem holds, so does the presumption of justifiedness. That is, even the social norm requiring the use of WhatsApp is justified to all who prefer coordination on some messenger service to no coordination. Any individual using a different messenger may then be criticized for failing to see that there already is an equilibrium solution in place that she can benefit from, instead of annoying everyone with insisting on her preferred messenger. Of course she may try to change the equilibrium in place or right out reject it as unacceptable – e.g. worse than having not coordinated on any messenger. But she nevertheless has to acknowledge that WhatsApp, as the already established equilibrium, is also attractive to her as long as she is in favor of coordinating on some messenger and knows that equilibria cannot be changed unilaterally.

Key Characteristics of Political Equilibrium

The WhatsApp example shows how some equilibrium norm can be uniquely justified in practice. Establishing such norms is a realistic and desirable open-ended ideal. It is realistic because it acknowledges that the norms we live by are usually not our most preferred norms. Therefore, they are (also) favored for pragmatic reasons, e.g. that they are the status quo, that norm

change is difficult, and that reasonable disagreement on the most preferred norms will persist. Having such norms is nevertheless desirable because it is at least better than having no norms in place at all and ideally is fairly close to the most preferred order.

I call this ideal "Political Equilibirium" (or PE). PE incorporates Gerald Gaus' idea of justified social order as some member(s) from a set of possibly stable and justifiable norms a given community might select. However, PE diverges from Gaus' theory in that it is strictly open-ended and procedural. Analogous to the idea of market equilibrium, PE provides a general and abstract understanding of a desirable social state. The purpose of this ideal is to guide us toward a social mechanism for bringing about such a state in actual society, not to derive universal rights or normative principles.

Let me summarize several defining characteristics of PE, resulting from the discussion so far:

Open-Endedness: We do not know for sure whether PE is possible for some given society. But, following Gaus, there are good reasons to assume that there is a set of justifiable norms for any given society.[30] Speaking of such a set of norms implies that we can be hopeful that justifiable equilibria do exist, whereas there is not one uniquely justified equilibrium from the perspective of normative theory. Accordingly, we cannot point to the correct equilibrium of any actual society. Thus PE is an open-ended ideal: It provides us with an abstract understanding of a desirable state in which people are governed by norms in coherence with their own preference for such norms. But it does not contain substantial claims about the correct justified norms of any given society.

Compromise: In line with Gaus observation that justified social norms are unlikely to be individually preferred norms, any instantiation of PE in large and pluralistic societies is probably a compromise.[31] A compromise is an arrangement where

> "[...] all parties regard some other arrangement – not the one agreed upon – as the optimal solution. Thus in a compromise, we have dissent on what would be the

[30] Again, it is of course not necessarily the case that there are. See also Fred D'Agostino (2013: 142 ff.).

[31] Gaus has some reservations about 'compromise' in the context of justification: "The deeper worry, though, is whether compromise is really at the heart of public justification. On the one hand, it may seem obvious that reasonable parties seeking to live together must exercise the virtue of meeting others halfway (i.e., splitting the difference). [...] However, when we take a broader view and understand public justification as deliberation about whether one's evaluative standards endorse a rule, the claim that the heart of the endeavor is about compromise looks dubious. To say that public justification involves splitting the difference between what a religious person believes is justified and what an ardent secularist holds to be supposes that living according to one's evaluative standards is like claiming a share of a common product, to be negotiated away." (G. Gaus 2011: 331-332) But surely this is just a disagreement on words. On my understanding, a compromise is not (necessarily) about "splitting the difference of a common product". It is rather about accepting "[...] that living with others involves accommodation to the fact that they have different standards, and we may have to accept that the justified rule is not the one we would have chosen if we were dictator." (G. Gaus 2011: 332)

> best arrangement, but we have consent that the arrangement agreed upon is better than having no arrangement at all."
>
> (Wendt 2016: 14)

In principle, it is of course possible that the social order we have and the one we prefer are identical, but this is highly unlikely. In a large pluralistic society we should expect that the norms we live by are not our favorite norms. This is not because someone has done us any injustice, but simply reflects the reasonable pluralism in the preferences there are. However, hopefully we can nevertheless see the order we have as an overall beneficial social tool of cooperation that provides us with at least enough reasons to prefer it to the state of nature.
Of course this leaves several important questions open. What, for instance, counts as the best possible kind of agreement a society in PE can hope to achieve? What about the space between the minimum and the optimum? And what kind of reasons correspond to the different kinds of agreement on an individual level? In order to answer these questions I will present a more precise account of justificatory agreement as a compromise in *Subsection 4.2.2* of the following chapter.

Proceduralism: In analogy to economic theory and the conception of market equilibrium, we should think about an open-ended ideal such as PE as being primarily procedural. That is, it is not about selecting the right outcomes – i.e. the correct norms – for some society in theory, but about using the ideal to guide us in thinking about the social process selecting outcomes. So process orientation means proceduralism, but with the procedure taking place in social reality, not in theory as is the case in HCM.
From the perspective of PE, and after having specified further what it is, an obvious way to proceed would be to again take a note from economic theory and start thinking about the right social mechanism for establishing PE. This however is an extremely complex task because it involves, firstly, extensive reflections on whether or to what extent democracy, the theory of democracy and social science have already attended this task. Secondly, it would further involve a collaborate effort of theorists, scientists and other experts to come up with recommendations for the best way of designing the PE mechanism, insofar as current democratic procedures fall short of its ideal. Needless to say, this is an immense task that I cannot possibly fulfill in the present work. I take some very first steps in the direction of thinking about the right mechanism of PE in *Section 5.1.*

Robustness: As Gaus argues, we want a system of norms that will tend to return to some equilibrium, but not necessarily the same equilibrium when disrupted; this is the difference between a system being robust and a system being stable. (G. Gaus 2016: 231)

As in the case of markets, in the political realm circumstances and preferences change. Thus, also the open-ended ideal of a PE should allow for dynamic adjustment. This however creates tension between the function of social order and the open-ended ideal of justified social order: From the ideal perspective it would make sense to demand a dynamic mechanism, such as a market, allowing for adjustments. On the other hand, there is the functional requirement that an instance of social order has to be stable in order to successfully coordinate our expectations and behavior. If we were to reinvent our political order every week, there would not be much order left to structure our interactions.

This is where we get into the *dis*analogy of economic and political equilibrium. In the case of economics, individuals can unilaterally change the overall distribution incrementally all the time by means of private decision-making. In the political context of common norms things are different. Norms and politics are – per definition – not things settled by private, but by public choice. Thus, individuals cannot unilaterally change equilibrium norms.[32] Now, within different systems of norm selection, e.g. different political systems, it may be more or less difficult to change norms. In any case, norms have to be somewhat stable in order to fulfill their purpose of coordination. At the same time an ideal such as PE must account for the possibility that norms change. So there is a tension in the idea of PE between the need for stability and the need for reform and adjustment. From a practical perspective this calls for a robust order that can be changed, even fundamentally, but not easily and perhaps in a piecemeal manner. This is precisely what constitutional democracies typically allow for.

Relativism: Some might be worried that a conception such as PE can only lead to all out relativism. That is to say that this kind of conception does not provide a critical perspective on social order at all, but rather proclaims that anything can be justified. I do not share this worry for two reasons. Firstly, because I believe relativism to be the proper default position in normative social theory – at least if confronted with similar problems as in this inquiry. Recall that the basic problem of justified social order under the assumption of reasonable pluralism as stated in *Subsection 1.2.2*: The theorist does not have access to people's well-considered preferences and even if she did, it might very well be the case that the given set of preferences would not lead to any conclusive outcomes. So right from the beginning, it does not seem that we have the resources to say anything substantial about good social order in theory. The calculation debate in economics is a case in point. Also, the basic idea of the social contract is not about the content of justificatory agreement, but about how such an agreement can be thought of in the first place. In order to make substantial claims about the content of the

[32] In terms of agreement, the difference between politics and markets is that a political order requires "conformity without unanimity", whereas market distribution produces "unanimity without conformity". (M. Friedman and R. Friedman 1980: 66)

contract, more (questionable) assumptions are needed.
Secondly, a procedural conception such as PE eventually becomes more and more substantial as we we think about what it means for an actual society to achieve it. To get an idea of how this thought process works, consider how the idea of market equilibrium fits into market theory more generally. Market equilibrium itself is just the open-ended ideal of having maximized consumer utility in a given economy. But when we think about how this ideal could be translated into a social process, a whole range of other ideas are necessary. In the case of economics they include competition, equality of opportunity, and a legal system. These ideas specify the conditions under which it is plausible to assume that an actual social process such as market exchange gravitates toward the ideal state of equilibrium.
This translation process of the abstract ideal into practice will however not follow a neat, deductive logic because the ideas that go beyond the open-ended ideal itself are not simply deducible from it. Rather they follow from practical considerations about the best way of putting the ideal into practice. And these considerations largely depend on knowledge, interpretation and analysis with respect to the social circumstances on the ground. Of course, reasonable opinions and expert recommendations may differ on such matters and thus also lead to different proposals on how to best realize the ideal. Therefore, the translation of PE into social reality will necessarily also employ inductive and abductive reasoning.[33] As always, more substantial claims come at a price. But in the case of PE I try to show that it is a price worth paying because, one, it shifts a lot of the discussions to the realm of social science where they can hopefully be handled more productively and, two, it leaves us with a more realistic and practically meaningful normative theory.
What we give up by proceeding in this way is the hope of arriving at universal principles of good social order in theory. But I do not believe this to be a great loss. Such principles are – as argued in the preceding chapter – difficult to construct and defend. In addition, such abstract principles lack practical meaning in that they require substantial interpretation and thus do not help us in deciding controversial cases. Generally, "[...] even if we actually had full confidence and complete agreement about the principles of justice, we would disagree about what social states best satisfied them." (G. Gaus 2016: 246) Therefore, abstract principles do not save us from difficult discussion about the best translation of the ideal into practice. Hence, I consider trading in universal principles of good social order for an open-ended ideal as an alleviation from unnecessary baggage, rather than a great loss.

As I pointed to at the end of the last chapter, I see a general tendency in the literature to abandon the quest for constructing universal principles of good social order in theory. This

[33] "Of course, like any claims about social realizations these may prove wrong, but that, I take it, is a benefit of, not a worry about, the analysis." (G. Gaus 2016: 176)

also implies a refocus on more open-ended perspectives. In Rawls the remaining image is that of a well-ordered society, structured by some reasonable and publicly justified constitution. (Freeman 2007: 255-256) Gaus on the other hand in his (2016) *The Tyranny of The Ideal* has turned to Karl Popper's notion of an open society. In Gaus' theory the open society is characterized by citizens' diverse perspectives on the matter of good social order.[34] This image seems to be perfectly suited to the notion of an open-ended ideal. But while Gaus makes a powerful case for the necessarily open-ended nature of the quest for good social order ("the ideal") over the course of the first three chapters of the book, he does remain reluctant to pursue an ideal that is also primarily procedural.[35] Overall, he seems to be convinced that there is much more to be gained by seeking further innovative models of the problem of justified social order, rather than thinking about an actual social mechanism. (G. Gaus 2018) I will continue to pursue the latter approach in accordance with the idea of an open-ended ideal as presented in this section. However, that is not to say that Gaus' approach of further modeling will not produce interesting results. As we will see in *Subsection 5.1.3* both avenues eventually prove complementary, rather than rival in coming up with a social mechanism of PE.

Adjusting the Second Research Question

This concludes our first glance at PE – i.e. justified social order as an open-ended ideal. But before we can move on, we need to adjust our second research question. Recall that in *Subsection 1.2.2* I stated the possibility question of social order:

(1) Under the circumstances of justification, how can social order be justified to each individual governed by its social norms?

and the *content* question of social order:

(2) What is the *content* of justified social order for some given community?

Now, the possibility question obviously remains as it is because it is just as relevant in the case of PE as it is in the case of HCM. The second question, however, has to be adjusted because, with the rejection of HCM in the previous chapter and the reflection on open-ended ideals in this chapter, it should be clear that I do not believe that a general answer to the content question can be given in theory. Rather, seeking an answer to the content question is precisely what we give up as we turn to the quest for an open-ended ideal. Thus I suggest the following replacement of the content question:

[34] "An Open Society, in which each is free to pursue his or her own inquiry into justice, exploring the terrain of justice as he or she sees it, using the methods he or she thinks most fit, will be characterized by continued, deep diversity, with no shared ideal." (G. Gaus 2016: 149)

[35] Gaus does consider some examples of how society can test alternative orders but renders them rather impractical. (G. Gaus 2016: II.4.1)

(2*) What is the practical meaning and relevance of an open-ended ideal such as Political Equilibrium for any give society?

Question (2*) challenges us to say something about how an open-ended ideal such as PE applies in practice. This is an important task, because open-ended ideals are, per definition, not practically meaningful – they provide us with a vague ideal but do not tell us what to do or what to change. In thinking about a remedy, two complementary strategies discussed in this section come to mind. First, there is the idea pursued by economics of devising a social mechanism (such as the market) for pursuing the open-ended ideal in practice. Second, there is Gaus' idea of a testing conception which allows us to evaluate social states respective the ideal. We will start discussing the latter strategy in the upcoming section of this chapter.

This concludes the second section of this chapter. The goal of this section was to come up with an ideal of justified social order which would allow us to dispense with hypothetical choice modeling. The idea I then pointed to was that of having an open-ended ideal such as market equilibrium in economic theory. Methodologically speaking, this idea consists in leaving the matter of good social order up to an ongoing and open-ended social process or mechanism. All that is done in normative theorizing is the formulation of what it would generally mean for a society to establish a justified order, irrespective of the specific norms it may consist of. The decisive difference between this approach and that of Rawls' and Gaus' construction of abstract principles consists in fully endorsing the open-ended and procedural nature of the ideal. In terms of outcomes this is the difference between constructing principles of good social order and constructing principles for selecting good social order. The former kind of principles makes substantial claims on outcomes, whereas the latter does not.[36] Accordingly, people can structure their society however they want, but they should do so for the right reasons.
This kind of approach accepts relativism as the default option. Therefore I have defended relativism as the appropriate starting point of normative theorizing, at least in the traditions of normative individualism and social contract theory. At the same time I ask readers who are put off by relativism to not disregard the entire project quite yet. Things will already become more substantial in the following section with the idea of agreement as participation.

[36]Rawls, for instance, beyond deriving his principles of justice, ends up defending. Gaus rejects it. More precisely, Rawls endorses a "liberal socialism", characterized by democratic government, free choice of occupation and market competition between worker controlled firms. (Rawls 2001: 138) As we have seen above, Gaus (2011: 511-521) argues for the "ineligibility" of socialism. The ideal of PE on the other hand can and must remain neutral on such matters.

3.3 A Testing Conception

In this section I seek a solution to a general problem with open-ended ideals captured by the modified second research question: If we do not have a substantial account of the ideal, how do we know where we are as actual citizens and societies relative to this ideal? How far away are we from utopia, in which direction does it lie and how would we even recognize it if we got there? These questions point to the requirement that any ideal theory must account for how it relates to social reality in order to be practically meaningful – i.e. in order to provide some practical orientation. Otherwise why do abstract normative social theory in the first place?
One solution to this problem provided by the example of market theory is to turn to the idea of a social mechanism for pursuing the ideal in practice. And thinking about markets we have at least an intuitive notion of how this can be done. We will start working on a more explicit account of a social mechanism of PE in *Section 5.1.*
The other strategy for adding meaning to an open-ended ideal such as PE is Gaus' idea of translating it into a testing conception. (G. Gaus 2016: I.2.2) However, I have rejected Gaus' way of doing so above because it builds on normative principles derived from a hypothetical choice model. What then could be a workable testing conception that does not rely on principles derived from HCM? This section is attempting to given an answer by explicating the practical idea of agreement as participation.

3.3.1 Agreement as a Test

Recall the basic logic of the social contract. The act of signing a contract in itself is neither necessary nor sufficient for judging some norm as justified. This is because the decisive underlying assumption is that of a norm that is justified relative to the well-considered preferences held by the respective individuals. The act of signing a contract and thus performing agreement is only of normative significance insofar as we take it to signal that the contracting parties have such preferences in favor of whatever they are agreeing to. More simply put, we believe that under the right conditions, someone signing – for instance a contract of purchase – *signals* that she freely endorses to make the transactions.[37] Otherwise, the act of signing the contract would be of no normative significance because then it could be the case that someone else simply forced her to perform the act of agreement. Thus, agreement is a signal and potentially a test: The act of agreement signals normatively significant individual endorsement, given conditions of free and well-considered choice. Asking for agreement under such conditions can thus provide

[37] "Thus understood the agreement is not itself a binding act – it is not a performative that somehow creates obligation – but is reason-revealing. If individuals are rational, what they agree to reflects the reasons they have." (D'Agostino, G. Gaus, and Thrasher 2017)

a test for determining whether some norm is justified or not.
Theories employing HCM have taken this test to be an act of hypothetical reasoning. That is, they are asking whether people *would* agree to some principle if they *were* to be asked for agreement under some ideal conditions, theoretically specified. The question of actual agreement on the other hand has been highly unpopular:

> "Certainly, no prominent theorist thinks that questions of justification are settled by an actual survey of attitudes towards existing social arrangements, and are not settled until such a survey has been carried out. The question, then, is not "Are these arrangements presently the object of an actual agreement among citizens?" (If this were the question, the answer would typically be "No".) The question, rather, is "Would these arrangements be the object of an agreement if citizens were surveyed?" Although both of the questions are, in some sense, susceptible to an empirical reading, only the latter is in play in present-day theorizing. The contract nowadays is always hypothetical in at least this first sense."
>
> (D'Agostino, G. Gaus, and Thrasher 2017)

There are several interesting things going on here. First, note that the first statement is plain wrong, at least if the range of "prominent theorists" is not constricted to social contract theory. A counterexample of a theory that takes actual surveys of attitudes to be of great importance is the work of David Miller, which we will briefly discuss below.
Note second that the worry with regard to actually asking citizens for agreement or disagreement seems to be that someone will probably voice disagreement. I am somewhat puzzled by this worry. If the idea was that the state hands out a questionnaire asking for people's approval of the present political order, promising to immediately dismantle said order given the disapproval of one citizen, disagreement would indeed be devastating. But this is not what we are talking about here. We are rather concerned with the question of whether, how, and perhaps *to what extent* our social orders could be justified. Intuitively, finding disagreement seems to be just as plausible and relevant as agreement. Should we indeed encounter disagreement, there are at least two possible reactions. One would be to suppose that the disagreement is genuine. That is to say that the act of disagreement shows said individual to have good reasons for rejecting the social order in question. In that case, the disagreement would pose a real challenge to the existing order and people caring for that order would be well-advised to learn more about the reasons for and extent of the disagreement. Eventually they will need a plan for dealing with this political rupture and potential source of disorder. The other possible reaction to disagreement would be to say that it is not genuine but due to some mistake. This

is probably the most likely motivation behind the worry of actual agreement and disagreement: Theorists who are so careful in constructing the most appropriate hypothetical choice situation are worried that a choice by an actual citizen is likely to be invalidated by non-ideal circumstances and faulty reasoning.[38]

In any case, since I already rejected the hypothetical approach in the preceding chapter, this kind of solution is off the table at this point. What is still on the table, however, is the option of sticking with the idea of actual agreement. If we could somehow make sense of this idea in light of the social world as we know it, this would be the kind of game changer we need for showing that the idea of PE can be practically meaningful. In more detail, if we had a plausible account of actual agreement, we could leave substantial questions of good social order and the question of the appropriate social mechanism aside. For then, we would have an evaluative tool for deciding whether, or rather to what extent, some actual social order conforms to the ideal of justified social order generally specified in theory. Having such a test would also be extremely helpful for the task of designing the appropriate PE mechanism. Firstly, because even the most ingenious mechanism may fail to deliver the desired outcome, having a mechanism aiming at producing justified social order *and* a test for the actual existence of such an order should always be seen as complementary. Secondly, if we had, as I will eventually propose in *Section 5.2*, a testing conception that provided an overall score of justifiedness, different mechanisms could be evaluated comparatively.

3.3.2 Surveying Agreement

One obvious way of approaching the matter of actual agreement is to simply ask people for their explicit opinions, using a survey. There are of course several ways of combining surveyed individual judgement and normative theorizing. Here I briefly discuss two variants: One, integrating individual judgments into substantial normative theorizing and two, surveys of institutional approval.

Surveys as Part of Substantial Theorizing

David Miller (2003) has made use of empirical insights into individual judgements on matters of just distribution in order to construct a coherent, pluralistic and empirically informed theory of social justice. The distinctive feature of this approach is that Miller constructs a pluralistic theory of justice in which citizens' judgments have the function of testing the validity of the principles constructed in theory for different social domains. I do neither wish to approve

[38] "I will argue that a Member of the Public is an idealization of some actual individual; a Member of the Public deliberates well and judges only on the relevant and intelligible values, reasons and concerns of the real agent she represents and always seeks to legislate impartially for all other Members of the Public." (G. Gaus 2011: 26)

nor object to Miller's theory here. His proposal is indeed of interest for those concerned with constructing theories of justice, including substantial normative principles. What I am suggesting in this chapter, however, is to leave aside the matter of justice and principles of justice altogether and rather focus on justified social order understood as an open-ended ideal. In this context, an approach that incorporates individual judgments into a theoretic construction of principles of good social order, is off the mark. Obviously, because an open-ended ideal is neither providing nor seeking such principles.

Surveys of Institutional Approval

What I do see as a more relevant approach are surveys asking for explicit approval of specific instances of social order. Consider as an example a survey of constitutional approval done by Nicholas Stephanopoulos and Mila Versteeg (2016). They asked 2215[39] Americans to state their approval of their federal and state constitution. They found an average approval score of 7.8 out of 10 for the federal constitution, while state constitutions earned an average rating of 6.7. In order to be able to explain these outcomes, the authors also included a range of questions about demographic attributes (gender, age, race, education, and income), civic knowledge (about the constitution specifically and current events generally), and institutional attitudes (toward one's state, country and party). Further, the survey included a range of policy questions in order to measure the fit between respondents' political preferences and the actual policies enshrined in the respective constitutions. Interestingly, it turned out that the fit between the individual political preferences and contents of the constitution did not have a significant effect on the approval rating. Rather, constitutional knowledge and "jurisdictional pride" turned out to be the most robust factors in explaining approval. The authors conclude:

> "The most important implication of our findings is that constitutional support cannot be won through constitutional refinement. Since neither charters' substantive content nor their non-substantive features influence approval, constitutional design is effectively useless as a tool for increasing public backing for the document. [...] Leaders who want their constituents to back their constitution are not powerless to bring about this outcome. But the right strategy is not to tweak the document to make it more attractive, but rather to boost people's familiarity with it and to swell their pride in their state or country."
>
> (Stephanopoulos and Versteeg 2016: 117)

The reason why I present this study here is that it nicely illustrates both the potential and difficulties that would face such a survey in a normative context. The main difficulty obviously

[39]The usable sample consisted of 2046 people.

lies with the fact that the individual reasons for approval or disapproval may be very different from the kind of reasons we are after in the context of an ideal theory of justified social order. To stick with the example of the presented study, it is questionable whether a kind of national pride can be a good reason for approving of a constitution. Even in principle, the experience of pride and having a reason seem to reside on different levels. Further, we know that national pride can be fostered by the very institutions – i.e. the state – that benefit from such supportive attitudes. In the US for instance, young school children are collectively pressured into "pledging allegiance" to the ensign. In general what we would prefer, from a normative perspective, would be approval or disapproval based on substantial knowledge and reflection upon the constitution in question. So there are obvious difficulties in coming up with a survey that can measure the kind of reasons that are relevant from the ideal perspective of justified social order.

On the bright side, the presented survey does also show that normatively relevant outcomes can be generated by such methods. In particular, the outcome that approval rates are positively correlated with constitutional knowledge seems to lend support to the hypothesis that the constitution in question enjoys significant degrees of support. The approval would be even more relevant from a normative perspective if one would further modify the survey so that the representative sample were only made up of individuals who can be expected to utter sufficiently well-informed approval or disapproval. Of course, individuals may still approve or disapprove for very different reasons. And some of these reasons may still be irrelevant from the perspective of ideal theory. Nevertheless, under conditions of well-considered choice, a strong tendency in favor or against some institution *on the aggregate level* seems highly relevant for matters of justification.

Still, there is the disturbing finding of Stephanopoulos and Versteeg that in their survey the coherence between the contents of state constitutions and individual preferences does not do any significant explanatory work. If this outcome should prove valid on a global scale, it would pose a fundamental challenge to efforts of measuring any actual agreement by means of surveys because if measured approval is independent of content, in other words, if people just do not care what is written in their constitutions, there seems to be no normatively meaningful agreement to be found at all. But there may be a plausible way to explain the findings by Stephanopoulos and Versteeg in a different way. To see this, first note that people may have some general knowledge about what a constitution is and what it generally includes (rights, procedures, political institutions etc.). Also, they may have a practical, day-to-day understanding of the kinds of basic freedoms and obligations that govern their lives (e.g.: "I can say whatever I want, as long as I don't insult people"). But it is entirely unrealistic to assume that they have any specific knowledge about their constitutional documents. Note

second that – following Gaus' idea of an optimal eligible set – we already suspect that there is a whole set of feasible constitutions that could be justifiable to some set of citizens. So we might conclude that Stephanopoulos and Versteeg have found that all US state constitutions are within a set of justifiable constitutions, feasible relative to a public of rather democratic and liberal individuals. All constitutions are, loosely speaking, good enough compromises for most Americans so that they would not worry too much about moving from one state to the other, even if constitutions differ in detail – details of which they are not very knowledgeable about anyway.

In conclusion, I do think that direct surveys of institutional approval can produce relevant outcomes for normative approaches in search for actual agreement to instances of social order. Especially if such surveys were specifically designed to this end and if we keep in mind that we need to be realistic about the kind of agreement or disagreement that can be expected from actual citizens. Perhaps the format of mini-publics – i.e. in-depth consultations of small, randomly drawn groups of citizens – would be more suitable than leaving people on their own with a questionnaire and the complex matter of constitutional choice to be dealt with within a few minutes.[40]
We should keep approaches that survey explicit agreement in mind, at least as a benchmark and test of validity for whatever else we come up with. For now, we turn my favored account of actual agreement: the idea of agreement as participation.

3.3.3 Agreement as Participation

The entry point to the conception of agreement I am seeking is the observance that the test of agreement does not necessarily have to be explicit (i.e. spoken or written). It is commonplace in everyday life as well as in social science to assume that people reveal their preferences through actions other than speech or writing. Accordingly, it has often been proposed that doing something or failing to do something reveals a preference or will for the consequence of the action – in that the agent may be thought of as *tacitly* agreeing or disagreeing to the respective consequence implied.

Toward a Helpful Account of Tacit Agreement

The idea I am after is already implicit in Thomas Hobbes' notion of a "tacit covenant". (Hobbes 1651: XVIII) But the classical accounts of tacit agreement are somewhat off the mark in the context of this inquiry because they were conceived in the context of consent as the basis for

[40] Although created from a different theoretical background and for a different purpose, James S. Fishkin's method of deliberative polling seems the most refined proposal in this area. (James S. Fishkin 1991: chap. 8)

political obligation[41] – a topic that does not concern us here. What is more, classic social contract theory does not offer a systematic account of tacit consent, only hinted at by Hobbes.

However, Craig L. Carr (1990) has already done us the favor of providing just that. Based on accounts by Grotius and Pufendorf, he proposes that:

> "Individual actions or general action plans will signal consent when they are embedded in a social context that identifies them as actions associated with participation in some rule-governed activity or association."
>
> (Carr 1990: 337)

This translates into the following more systematic account:

> "Anyone who does X (where X = a conventionally understood indication of participation in some rule-governed activity, association, or enterprise P) signals his participation in P, and thereby consents to obey R, where R = the rule system constitutive and regulative of P. Consent here is expressed tacitly as a logical consequence of the decision to do X and participate in P."
>
> (Carr 1990: 337)

To illustrate the mechanism, Carr points to examples of participating in games. If, for instance, someone sits down to play a game of chess or walks onto a tennis court to play tennis, she is, insofar we take her to indeed engage in a meaningful social practice, logically committed to consenting to the rules of said game. For the act of playing chess or tennis is only meaningful, if one understands the rules constituting such games.
In order to avoid obvious objections against the cited definition we need to assume that one, the act of partaking is done voluntarily and two, that it is done in knowledge of what said social practice is generally take to consist of. This further implies that what actually counts as having consented is conventional and context dependent. (Carr 1990: 338; 342)

Agreement as Participation and PE

I believe the notion of agreement as participation is just the kind of idea that we need in order to construct an open-ended theory of justified social order. The main reason being that tacit agreement as participation is a straightforward feature of our everyday social interactions, which goes well with an account that maintains good social order to be determined through actual social processes. The link between the two ideas is that voluntary and well-informed

[41] John Locke (1690: §119) is another example.

participation can signal agreement to the social norms that are known to be part of some practice. This agreement can in turn be taken as an indicator that participants have their own good reasons for endorsing the norms that persist as part of this practice.

As Craig Carr suggests, a pertinent example of this would be partaking in a familiar game such as chess. Under normal circumstances, we would expect that two players sitting down at a table, opposite each other, setting up the board and initiating the game by the first move of one player, know exactly what is implied by the social practice they are voluntarily engaging in. Thus, under normal circumstances, observing two people doing this is sufficient for coming to the conclusion that they are doing so for their own good reasons. Consequently both chess players are now bound by the rules of chess. And if one of them were to break the rules, say by moving the king two squares, the other would be justified in criticizing the other for obviously violating the well-known rules of the game they are playing.

One thing we notice right away is that the agreement inferred from the act of participation is a special kind of agreement in that it does not directly relate to the rules we are seeking to justify. More precisely, the players, in setting up the game and starting to play, primarily signal that they agree to playing chess and not directly to any specific rule of the game, such as the rule that the king may only be moved one square at a time. The act of participation primarily signals that players value the practice as a whole and thus can be said to have their own good reasons to engage in it. The rules are of course an essential part of this practice but they do not exhaust all that it consists in. Therefore, acts of voluntary participation in some practice do not tell us much about the relation between individual reasoning and particular rules. Hence, a given chess player may not particularly like or may even outright hate some particular rule such as the possibility of castling. In a private game, she therefore might always start an argument with the other player on whether they should change the game by removing the castling rule. In an official tournament, however, she always has to accept this detested rule as a, for now, fixed feature of the game she loves to play. What this shows is that agreement as participation is often a rather indirect, pragmatic kind of agreement to a package of norms in the context of an overall valued practice. I believe this kind of agreement links up quite well with the idea of PE and justified social norms that are often not people's preferred norms.

Getting back to the matter of justifying social norms, we can see that the test of agreement as participation may not be well-suited for all norms and all situations. This is because, for the test to be applicable, we need norms that are embedded in voluntary social practices. Again, there are pertinent examples of this, such as norms that are involved in going to a restaurant. Suppose for instance that there is a Chinese restaurant that only allows its customers to eat with chopsticks. Clearly, going to any restaurant is a voluntary act and there are usually several restaurants to choose from. So any customer ordering a meal at a Chinese restaurant

can be said to have agreed to the "Chopsticks only!" norm, given that she is familiar with the norm of said establishment when doing so.

But things get more difficult when a norm is not neatly embedded in a voluntary practice. Think for instance of norms that apply to society at large such as a greeting norm. Such a norm might state that every person ought to greet any other individual they know upon meeting them in public. Because the scope of this norm is so broad, the respective practice it is embedded in is basically all of social life. And since partaking in this practice is usually a necessity for every member, actually observing someone doing it cannot be taken as a meaningful agreement to the rules of this practice. We can of course think of a scenario where the test does apply. For instance, it could be the case that people can voluntarily participate in sub-group activities, say different clubs where different kinds of greeting norms exist. Then, if a person were to join some club governed by the firm handshake norm, this could be taken as an agreement to the firm handshake norm, assuming that joining clubs is voluntary and that, ideally, there are different clubs with different greeting norms to choose from.

Besides norms with a society-wide scope, further problematic norms for our testing conception of agreement as participation are norms that themselves require participation. Think again of our WhatsApp example from above. WhatsApp, so our story went, is a conventional equilibrium that turned into a social norm. That is, once most people coordinated on using WhatsApp for prompt communication, single individuals refusing to use it became an annoyance, giving rise to the widespread normative expectation that people should use WhatsApp. Now, as soon as this transition from a mere convention to a social norm has taken place, our test of agreement as participation, by definition, no longer applies. This is because the norm in question itself demands participation, whereas our test requires conditions of voluntary participation. In such a case we would have to take a step back and ask whether we can describe the social norm demanding participation as being embedded in a broader, voluntary social practice. Perhaps in this case it is plausible to argue that using WhatsApp is partaking in the broader practice of digital communication and partaking in this practice is voluntary, because one could also stick to more traditional means of distanced communication, such as calling people on the phone are writing e-mails. So as long as the "Use WhatsApp!" norm is only understood as applying to people partaking in the practice of digital communication, the test could still be applicable. If, however, the "Use WhatsApp!" norm is understood as a norm that applies to everybody, we are back to the problem of identifying a practice that can still be reasonably interpreted as voluntary.

The overall lesson from these examples is that the testing conception of agreement as participation may not be equally applicable to all norms in all scenarios. But this is not necessarily a problem for our project of testing the justifiedness of social order. It should not be a problem

if we can indeed identify instances of voluntary participation in respect to the social order in question. This brings us to the case of political participation.

Agreement as Political Participation

Leaving the discussion of simple examples aside, the core question for our testing conception of PE then becomes: What would be the relevant act of participation and what would warrant the conditions of voluntary and well-informed action in the case of justified social order? In the context of PE, an obvious answer consists in pointing to instances of political participation. Participation in politics, so the idea, can – under the right conditions – be interpreted as agreement to the rules of this social practice: As a player engaging in a game of chess can normally be understood as agreeing to the rules of chess, a citizen engaging in the, as it were, game of politics can be understood as agreeing to the rules of the respective political system. Elaborating this idea of agreement as political participation will be the core goal of *Section 5.2*. In more detail, the task will be to identify acts of participation and the real-world conditions for voluntary and well-informed action that could qualify as the actual test. Further I will suggest that, once we have identified these parameters, they can be translated into an index of justified social order, much like a democracy index. Doing so has several advantages. One, it allows us to use empirically methods in order to determine the justifiedness of some existing instance of social order. Two, operationalizing the PE in this way translates it into something that has a very precise practical meaning. Three, having an index as a testing conception allows us to work with established methods and data from the field of democracy research instead of having to come up with an entirely new test.

This concludes the third and final section of this chapter. The goal of this section was to find a viable testing conception for PE that does not rely on HCM. Ultimately I have tried to show that actual agreement in the form of voluntary and well-informed participation can be understood as such a test. Hence, a normative theory with an open-ended ideal in combination with a testing conception of agreement as participation would be a significant step toward a practically meaningful theory of justified social order.

What remains to be specified are the details of our open ended ideal, how actual social states can effectively be evaluated respective this ideal and how we can evaluate non-existent social orders and come up with reform agendas. Some answers will be provided in the following two chapters.

3.4 Concluding Remarks *Chapter 3*

Overall, this third chapter proposes three building blocks of an alternative, procedural theory of justified social order: One, a descriptive account of social order understood as a set of social norms, two, PE as an open-ended and procedural ideal, and three, a testing conception of agreement as participation. These ideas will eventually be molded into one coherent theory of Political Equilibrium in the following two chapters. In this chapter, the following has been argued:

1) Proponents of hypothetical choice modeling usually base their theorizing on theoretic and normative conceptions of social order. These conceptions have an unclear relation to social reality, motivate endless debates and the assumption of a questionable normative sphere of morality or justice.

2) The remedy I propose is to put conceptions of justice and morality aside and start out with a descriptive or empirical account of social order. My hope is that doing so will allow us to leave debates on the best conceptions of social order up to the empirical sciences and produce more convergence in theorizing.

3) So far I consider Cristina Bicchieri's account of social norms as the most suitable empirical account for a general theory of justified social order. Most importantly, social norms explain why patterns of norm following occur and how the perspective of justified norms lets us intuitively distinguish between good and repressive norms.

4) In order to further develop the normative perspective of justified norms without the use of hypothetical choice modeling, I have turned to the idea of an open-ended ideal. Taking notes from economic theory, an open-ended ideal describes a general desirable state to the end of guiding a social process for establishing justified outcomes, *without* making claims about the right outcomes.

5) With Gerald Gaus' idea of an optimal eligible set of possible justified equilibria, we get a minimalist model view of an open-ended ideal for the political realm: An existing status-quo norm can be uniquely justified as the already established device of coordination and cooperation in some society, given that it is at least preferred by all members to not having any norm in place at all.

6) In order to explain how an open-ended ideal can be practically meaningful without returning to the abandoned idea of hypothetical agreement, one obvious move is to start thinking about the right social mechanism for establishing justified norms. Another, less

demanding idea is that of a testing conception for the existence of justified order in a given society.

7) A good candidate for the latter is the idea of agreement as participation. This draws on the everyday observation that people participating in some social practice can be interpreted to agree to the rules of said practice under conditions of voluntary and well-informed action. This allow us to interpret political participation as agreement to the rules of the respective political system.

8) With this testing conception we get something that could eventually be translated into an empirical test such as a democracy index. This would allow a systematic test and – to some extent – a comparison of existing social orders in terms of their justifiedness. Such a test would clarify what it means for some actual society to establish justified social order and to what extent it has already done so.

The Broader Message of Embedded Constructivism

As stated at the beginning, there are two sides to the narrative in this chapter. One side is about the ideas we need in order to construct a theory of justified social order without HCM. Each of the three sections in this chapter explores one such idea. The other side of the narrative is about carving out the contours of an alternative methodological approach to HCM. I call this approach "Embedded Constructivism". Here, I explain how it works.

Obviously, Embedded Constructivism is a kind of constructivism. As such it involves some building material (i.e. some basic assumptions) and a procedure for arriving at an answer to, in our case, the question of justified social order. In the previous chapter we discussed Rawls' hypothetical constructivism. This type of constructivism is characterized by building on a normative conception of social order (the well-ordered society) and a hypothetical procedure (choice in the original position) for deriving substantial normative claims (the two principles of justice).

Embedded Constructivism differs from Rawls' hypothetical constructivism in several respects. Firstly, it starts out with a different kind of building material, namely with a descriptive account of social order – of people and their norms. These things are explicitly laid out in what I call the "empirical model" in the upcoming *Section 4.1.* As in Rawls' hypothetical constructivism, Embedded Constructivism also needs normative building material in order to get to a normative perspective on social order. In Rawls' case this is the idea of justice as fairness. In my case it is the JPN. I suppose that in both cases the normative building material is a *re*construction of existing normative thought. Simply put, the normative building material is something people allegedly already care about deeply.

Secondly, embedded and hypothetical constructivism differ in the status and functioning of the respective procedure. In hypothetical constructivism the procedure is a theoretical fiction. In Embedded Constructivism the procedure is an actual social process, and providing a guiding ideal for this process is the central outcome of the construction.
Thirdly, embedded and hypothetical constructivism differ in their outcomes. Hypothetical constructivism deduces substantial normative claims of justified social order. Embedded Constructivism translates the open-ended ideal into practical devices (i.e. a social mechanism or a testing conception) that actual citizens can use to pursue the ideal specified in theory. That is, it systematically relates the outcomes of theoretic normative constructions back to social reality.
This third aspect completes the "embeddement"; an enclosure of the normative construction, or "normative model", between two things. One, a descriptive account of people, the orders that structure their social lives and what they value about them – "the empirical model". Two, a practical account of how the ideal can be pursued in social reality.
The image that emerges is that Embedded Constructivism works much like modeling in the social science. The empirical model in Embedded Constructivism models aspects of social reality such as people, social order and what they value about it. The normative model reconstructs an abstract ideal based on assumptions about what people actually value (much like HCM according to the empirical reading). In doing so the descriptive account works as an anchor, ensuring that the normative model – although describing an ideal state of affairs – remains a model about actual people, their problems and reasons. In terms of the outcomes, modeling in the social sciences produces predictions about what people will do. If the predictions turn out to be wrong, the model does not fit the facts. Normative modeling in Embedded Constructivism produces predictions about what ideals and practical devices people have good reasons to endorse. If they reject the outcomes of the model even on due reflection, the model does not fit its protagonists. This is why the aspect of self-testing, expressed by *Condition 3* of the JPN, is so important.

None of the above is meant to say that existing theories do not already embody or even defend Embedded Constructivism. My criticism is rather, especially in respect to Rawls, that while his theory is explicitly about our actual social orders and providing some guidance in assessing them,[42] the embedding is not systematic. Rawls, instead of relying on much (social) science, bases his construction mainly on armchair sociology and never explains how we will find out whether full justification and full publicity can be attained by his or any other proposal.[43]

[42] See his *Four Roles of Political Philosophy* in Rawls (2001: 1-5).
[43] Analogously, Gerald Gaus fails to tell us how we, as citizens, can establish our optimal eligible set and choose from it.

Therefore, his account of a well-ordered society and its principles of justice seems to be floating above, instead of being embedded in, social reality.

The sketch of a theory of justified social order I will lay out in the following chapter is an exemplification of how Embedded Constructivism can be employed. My hope is that even readers who disagree with (many) aspects of this theory will nevertheless be convinced that the broader message of Embedded Constructivism is worth considering.

Chapter 4

A Theory of Political Equilibrium

In *Chapter 3* I suggested two kinds of antidotes against the shortcomings of hypothetical choice modeling (HCM). First, a general account of how to profess normative social theory, which I refer to as "Embedded Constructivism" that aims for a systematic embedding of normative theorizing in social reality on both ends. More precisely, I have argued that normative theorizing should start out from a descriptive conception of the object of theorizing and ultimately relate back to social reality. Second, I have suggested three ideas or "building blocks" that allow us to live up to the demands of Embedded Constructivism while theorizing about the problem of justified social order: social norms, an open-ended ideal called Political Equilibrium and agreement as participation. In this chapter, I combine the first two building blocks toward sketching a *theory of Political Equilibrium*.

The goal of this Chapter is to specify in more detail what Political Equilibrium (PE) is all about and, in doing so, exemplifying how Embedded Constructivism may be applied. At least in respect to the two steps of laying out a descriptive account of social order – the "empirical model" developed in *Section 4.1* and constructing a well-suited "normative model", explicating PE in *Section 4.2*. The third step of Embedded Constructivism – relating the outcomes of the normative model back to social reality – will be attended to in *Chapter 5*.

4.1 The Empirical Model

In *Chapter 3* I have argued for starting out with a descriptive account of social order, because this increases our chances of converging on a common reference point in theorizing, hopefully facilitating more convergence in normative social theory.

The preliminary descriptive accounts sketched in *Chapter 1* and *Chapter 3* depict social order as a set of social norms, governing social interactions of some set of individuals. Social norms in turn are to be understood as behavioral equilibria, supported by symmetric beliefs and

expectations. This means that some rule prescribing that one must do or not do X exits in the shape of coherent empirical and normative expectations in support of X, held by most individuals, while at least some of them are ready to also sanction transgression. But how come we have such a remarkable social tool?

In order to answer this question and thereby learn more about the object of our inquiry, we must refer to what I call the emerging theory of social order. In this first section I provide an overview over this theory, which falls into two main parts: one, attempts to explain cooperation, and two, approaches to explaining coordination.

The core lesson from this summary is that human beings are at least as much community-oriented cooperators as they are self-oriented utility maximizers. As such beings they possess the ability to coordinate on cooperative norms. This insight simultaneously offers a common starting point for normative theorizing and has important implications for what would be an appropriate approach to a problem such as justifying social order. More precisely, it implies that a normative account probably cannot tell us what specific norms some society should select, but perhaps it can offer some guidance on how to design the process of selection.

4.1.1 Cooperation: Solving Collective Action Problems

Two things are certain: cooperative social order is highly beneficial and human beings are – besides other things – a highly cooperative species. Nevertheless, rational choice theory tells us that there are fundamental problems of cooperation: so-called collective action problems.

The classic illustration of this problem is the "prisoner's dilemma". This game theoretic abstraction from actual social situations constructs a dilemma between the choice of a cooperative and an uncooperative option. The dilemma consists in the tension between the long-term benefit of cooperation and the short-term benefit of free riding and the problem of uncertainty. A typical real-world example for this problem would be the choice between buying a ticket for public transportation or dodging the fare. Today you might save some money by dodging the fare, but note that in the long run, many people failing to pay will cause ticket prices to go up for everybody and this might eventually cause the whole system to break down. So you might think that buying a ticket – the cooperative option – must be the right thing to do, but then you realize that you also cannot be sure what others think and do. Most others might actually dodge the fair and then you would be the "sucker" who buys a ticket and thereby also pays for the free riders.

If we had a social norm in place, effectively prescribing people to always buy a ticket, the cooperation problem would be solved. To have such a social norm in place would mean that most people prefer to buy a ticket because they believe that others will also buy a ticket and expect them to do the same. Whereby the normative expectation to buy a ticket might be

further supported by the belief of looming formal or informal sanctions.

Toward a Better Model

The first important point here is that within the framework of classic rational choice theory, we cannot reason ourselves out of the prisoner's dilemma and into the cooperative world of social norms. Therefore, the "rational fools"[1] always remain stuck with non-cooperation. Thus, in light of the fact that we do achieve cooperation, there must be something rational choice theory is missing.

Fortunately, biology, psychology, behavioral economics, philosophy and in particular the tool of evolutionary game theory[2] have already provided most of the missing parts of the puzzle for explaining stable cooperative social order.

Gaining Insights From Evolutionary Game Theory

Evolutionary game theory has produced a whole range of fascinating results, relevant to explaining the emergence of cooperative social order. The first important insight stems from Maynard Smith and George Price (1973) and concerns the solution concept of evolutionary game theory. In their seminal paper they provided us with the concept of an evolutionary stable strategy (ESS). The concept of ESS teaches us that in an evolutionary context a stable strategy does not only have to outperform all competitors in a given population, but it also has to outperform all new ("mutant") strategies which might invade the population in the future. Further, there are four key insights for understanding the human ability to achieve stable and cooperative social order.

The first piece of the puzzle is *reciprocity*. We know from the work of Robert Axelrod (1984) that more complex kinds of cooperative strategies can in principle outperform free riders. These more complex strategies are so-called *mixed strategies*, which combine different responses. In particular, Axelrod has shown that strategies practicing reciprocity are a key component of producing stable cooperative states. They do so by punishing free riders with non-cooperation, while cooperating with other cooperators.

The second piece is *correlation*. Correlation between similar strategies increases the likelihood of stable cooperative order. You can think of this as people interacting more often with

[1] Amartya Sen (1977) famously criticized the purely instrumentally rational concept of human beings for misrepresenting them as rational fools.

[2] In evolutionary game theory players still play simple games, however they do so repeatedly while being paired at random with other players of the same populations again and again. Thereby the players are no longer assumed to be rational in any sense. Rather, players are now really nothing other than strategies. Further, through simple models of learning by imitation – e.g. the *replicator dynamics* – strategies change dynamically. That is, their distribution in the population changes according to how well they do compared to how well the other strategies do. It is also possible that more complex strategies – so called *mixed strategies* – emerge and that new or mutated strategies can invade a population.

relatives, friends, or neighbors than with people they do not know. A biological reason for this correlation could be kin selection as described by William Hamilton (1963). A social reason for correlation could for instance be reputation effects, which enable "indirect reciprocity", i.e. the phenomenon that two strangers cooperate as if they had cooperated successfully before.[3] Generally, correlation has a positive effect on cooperation because cooperative strategies do well against themselves, while free riders do not. Hence moderate levels of "positive correlation of strategies with themselves is favorable to the development of cooperation and efficiency." (Skyrms 1996: 61-62)

The third piece is *strong reciprocity*. This idea is based on experimental findings in cooperative games - i.e. public good games - which show two interesting phenomena: One of them is that there always seems to be a mix of egoists, who tend not to contribute, *and* cooperators, who regularly do contribute significant amounts to a mutually beneficial public good. The second interesting phenomenon is that there are some individuals willing to spend some of their resources on punishing free riders. This enforcement behavior leads to significantly higher levels of overall contributions to the public good. (Gintis 2008) Ernst Fehr and Urs Fischbacher (2004) further show that even unaffected third-party agents often bear the costs of punishment. Strong reciprocity also seems to demarcate the boundary between mere biological evolution and what may be described as biological-cultural co-evolution. This is because, as the example of vervet monkeys shows, reciprocal altruism predates human animals. (Cheney and Seyfarth 1990; Skyrms 1996) Whereas Samuel Bowles and Herbert Gintis' (2004) work on strong reciprocity suggests that early humans could only achieve strong reciprocity due to greater cognitive and linguistic capability relative to monkeys. Hence, cooperation in large anonymous groups is most likely a cultural refinement of strong reciprocity, unique to humans. (Bowles and Gintis 2011)

The broader image that emerges out of these insights is that a stable cooperative order is made up of rule followers, a small amount of rule breakers, and rule following punishers, who punish the rule breakers. (G. Gaus 2011: III.7; Ostrom 2000) Evolutionary models of strong reciprocity as in Bowles and Gintis (2004) show that such an overall cooperative order of norms can be evolutionary stable. And "[w]hile no full-blown theory of collective action yet exists, evolutionary theories appear most able to explain the diverse findings from the lab and the field and to carry the nucleus of an overarching theory." (Ostrom 2000)

An Emerging Theory of Cooperative Order

This brings us to the fourth and last piece of the puzzle: social norms. As stated in *Chapter 1*,

[3]For a model of indirect reciprocity see Robert Sugden (1986). For a contemporary study of the effect of reputation, illustrating indirect reciprocity, see Andreas Diekmann et al. (2014).

social norms are behavioral equilibria that exist in the form of a coherent web of empirical and normative expectation regarding what is done in some situation, plus a sanctioning mechanism. And their existence makes a lot of sense in populations where many people follow the rules, while some do not and thus need to be policed by those ready to punish rule breakers at their own expense.

Generally, the symbiotic relationship between the literature on the evolution of cooperative order and social norms consists, on the one hand, in the former explaining the possibility of the latter. For instance, the insight that strong reciprocity can be evolutionary stable explains how collective action is possible at all in large populations. On the other hand, social norms provide an overarching theoretical-empirical framework that can explain how rule following is part of our everyday practices and our psychology, how different conceptions of human nature (e.g. egoism and altruism) fit together and how this helps us to make sense of rule following as well as rule breaking. Elaborate accounts of social norms such as Christina Bicchieri (2006) do all of these things. Thus, there is not much left standing in the way of a full-blown theory of cooperative social order. A theory that might eventually turn into a viable overarching theory of social science, replacing rational choice theory, which "has produced significant insights. But it may have run its course." (Bowles and Gintis 2011)

Now, some caution is advisable, because although the models and experiments cited above do provide some explanations for the possibility and nature of cooperative social order, these are mere abductions constructed from what we know of biological evolution and patterns in human behavior that we can observe today. So for example, "[w]e do not know that a human predisposition to strong reciprocity evolved as we have described. But our simulations suggest that it could have." (Bowles and Gintis 2004: 27) In absence of a better explanation, however, we should consider these kinds of abductions a crucial first step toward a complete scientific theory of evolved social order.

Also, note that all of these insights reside on a fairly high level of abstraction. As a consequence, they do not explain any particular order – the particular set of norms – we actually encounter in our social lives. The emerging theory of cooperative order can explain how it is possible that we have such norms and how the capacity to have them might have evolved long ago, but this story leaves out a lot of factors that are important. Namely, things such as history, culture, and power relations that have shaped the norms we have.

Generally speaking, we have some answers to the question as to whether and how collective action problems can be solved to the benefit of cooperative social order. But having overcome the problem of cooperation we notice that there is another tremendous challenge: establishing one particular cooperative social order that does the job of coordinating our expectations and

behaviors in our daily lives. Looking at human societies – past and present – almost infinite different sets of norms seem to be feasible in principle. So how to precede? We can start by considering the literature on coordination, which at least gives us a better understanding of the underlying problem and some glimpse at possible solutions.

4.1.2 Coordination: Selecting Among Several Alternatives

In game theory, coordination problems are situations in which individuals need to coordinate on one of at least two feasible solutions. A standard example would be two people who try to meet, say, at the airport, but did not specify exactly were to meet. The important underlying assumption of coordination problems is that the individuals involved strictly prefer coordinating on any of the available solutions to not coordinating at all. Further, a distinction is made between pure and impure coordination problems. In a pure coordination problem, all individuals value the different feasible solutions equally. In an impure coordination problem, different individuals prefer different solutions. In the airport example, it could for instance be the case that one person prefers meeting at a coffee shop, whereas the other prefers to meet at a restaurant. This would be an impure coordination game, because while both are still assumed to prefer any meeting point to not meeting at all, there is also a conflict of interest involved.

I take the selection of social norms to be an impure coordination problem. This follows from two considerations. The first one is that all feasible norms are also solving a cooperation problem that has the structure of a prisoner's dilemma – we are, by definition, within some problem of coordination. To see this, note that if a social norm solves such a cooperation dilemma, the outcome is per definition preferable to having no social norm in place at all. Having to choose between different options which are all at least preferable to not having any option realized, is the defining characteristic of a coordination problem. The second consideration is that a pluralistic society, characterized by disagreement about the right norms, is best modeled as an impure coordination problem. Modeling it as pure coordination problem would imply that pluralism is only ignorance of the one solution that everybody does in fact prefer.

How Do We Solve Coordination Problems?

Now, given that social norms solve impure coordination games, what do we know about solving such problems? Why do we coordinate on one set of norms rather than another? Why, for example, do the British drive on the left side of the road, while other Europeans drive on the right side? We do not have anything close to a complete theory of norm selection. Nevertheless, there are some interesting partial answers to consider.

One of the first partial answers is provided by David Lewis (1969). Lewis discusses conventions

only as solutions to coordination problems. Conventions, according to Lewis, work in the same way as social norms, minus normative expectations and sanctions. That is, conventions coordinate behavior by means of symmetric conditional preferences for doing something, if one expects that the others will do the same. Hence, in a meeting game, such as the above-mentioned problem of meeting at an airport, "I may go to a certain place because I expect you to go there, while you go there because you expect me to". (Lewis 1969: 25)

So how do we coordinate our expectations? Lewis offers two answers: agreement and salience. Agreement denotes the obvious fact that we can coordinate expectations by talking about what we will do. Coordinating by explicit communication also has the advantage of creating higher order expectations. This denotes the idea that if two people A and B discuss and agree to meet at Joey's restaurant tomorrow, they will both form the first order expectation that the other will go to Joey's tomorrow. Since A has explicitly told B that she will go to Joey's tomorrow, she will probably also form the second order expectation that she expects B to expect her to go to Joey's tomorrow. Understanding this, B might form the third order expectation that he expects her to expect him to expect her to go to Joey's tomorrow and so on. Lewis' claim is that higher order expectations increase the likelihood and stability of successful conventions.

Although communication resulting in explicit agreement is a very effective way of establishing conventions, Lewis is more interested in implicit ways of coordinating expectations.[4] Hence, also drawing on Thomas Schelling's concept of a focal point[5], Lewis uses the idea of salience. "Salience in general is uniqueness of a coordination equilibrium in a preeminently conspicuous respect." (Lewis 1969: 38) For a philosopher such as Lewis, known for his formal methodology and rigor, this is quite an imprecise notion. His most used example of what it could mean is precedent. That is, it could be the case that for some reason people can draw on their past experience when facing a coordination problem and find a solution that accords with something that has already worked before. Besides precedent, however, there could be many and entirely arbitrary reasons for one of the possible coordination equilibria to become salient.[6] Essentially, salience is simply the fact that *for some reason* an option might stand out and thus becomes the obvious solution for the coordinators.

[4]This is because his ultimate aim is to explain how we can coordinate on a language without presupposing that we already have some language in virtue of which we achieve this.

[5]See my summary of Kaushik Basu (2015) in *Subsection 3.1.3*.

[6]Assume for instance that you and a colleague of yours agreed to meet at a restaurant in the city center yesterday, but failed to specify the exact restaurant. Further assume that you have no way of contacting your colleague beforehand and your meeting starts in 20 minutes. Where do you go? Well, what if you were to remember that yesterday the both of you happened to have a lively discussion about the benefits of being a vegetarian. Remembering this and also expecting that your colleague might remember it, you start looking for vegetarian restaurants and as it turns out there is only one in town. Thinking that your friend probably will end up at the same conclusion, you go to that vegetarian restaurant. In this case some arbitrary fact - you and your colleague having a conversation about some topic - makes one option salient.

Despite the imprecise concept of salience, Lewis' analysis shows at least three things. One, there are usually several feasible equilibria to coordinate on. Two, there are several ways of coordinating on one of them - an obviously good way would be communication, but there are also more tacit ways of establishing conventions. Three, in cases where conventions do not arise intentionally through explicit communication, we do not really know what happens. We only know that for *some* reason one option has to be made salient. Interestingly, experiments show that we are highly sensitive to cues of salience, especially in the behavior of others. (Cialdini, Reno, and Kallgren 1990; Faillo, Grieco, and Zarri 2013)

Brian Skyrms (1996) criticizes Lewis for assuming too much common knowledge as a precondition for conventions to arise. He replaces the idea of a salient equilibrium with that of a *correlated* equilibrium. Skyrms argues that we should think about the coordination problem as a learning problem in an evolutionary model. His claim is that in such a setting people can learn to play an equilibrium by correlating their choice with some random external event. Here is Skyrms' real world example:

> "When two motorists meet going opposite directions at an intersection, one sees the other on her right, and the latter sees the former on her left. As far as the motorists are concerned, being on the right or the left is a random event. One correlated equilibrium is "the rule of the right"; the driver on the right goes first. This norm actually did evolve. The alternative "rule of the left" is another, perfectly acceptable, correlated equilibrium that did not evolve."
>
> (Skyrms 1996: 75-76)

Brian Skyrms and Peter Vanderschraaf further argue that learning to play a correlated equilibrium can be understood as "inductive deliberation". Here the idea is that players correlate through some process of inductive learning. That is, they form beliefs over what other players will do and update these beliefs based on the information they gain by each interaction. Introducing such simple learning dynamics into the model can show how individuals can establish a convention over time. (Vanderschraaf and Skyrms 1993, 2003) So at least such models offer an explanation as to how it is even possible that a group of players learns to coordinate on a norm.

Social Norms, Again

Overall, (evolutionary) game theory does provide some interesting insights into how impure cooperation problems can be solved. But since these models typically assume the absence of any social fabric or communication, the game theoretic approach to problems of coordination remain highly abstract and incomplete. (Bowles and Gintis 2011: 5.5)

Once again, social norms seem to be an important conception for getting a more complete picture that relates to everyday practices of norm following. That is, if we assume the existence of a social norm in the case of some "game" – i.e. some social scenario such as the meeting of two motorists at an intersection – we get a more straightforward understanding of how people "inductively learn" to play "correlated equilibria". For, if we assume that social norms exist, it is quite obvious that people can coordinate on an equilibrium by learning from conversation or observation what kind of norms are commonly at play in a given practice, such as using public roads.

Perhaps also game theoretic modeling will become less incomplete as more and more core human capabilities are included into the model. The introduction of simple learning dynamics by Sykrims and Vanderschraaf is a case in point. Another obvious candidate would be conversation.

Perhaps the most peculiar aspect of game theory is that players usually cannot do what we do every day in our actual social lives: We talk about things. Within the above-cited research project of solving coordination problems, this peculiarity is due to philosophical obsession with explaining the emergence of language itself. (Vanderschraaf 1995: 82) And although we might indeed learn something valuable by explaining language from a pre-linguistic stage, we are also sure to miss a great deal if we confine ourselves to such explanations of social order.

4.1.3 Communication, Reasoning and Norms

As actual social beings, we discuss and debate the norms we live by. In fact, the social practice of explicitly discussing and creating norms by means of language – i.e. politics and similar activities – makes up a significant portion of what one might call our social world. It would thus be strange if the use of language were not an important part of understanding social order. Also, a theory of justifying social order – a theory of reasoning about social order – presupposes that especially reasons and reasoning actually matter for how social order works. So here are some core insights from the empirical literature on the interplay between reasoning and social norms.

The Interdependence of Language and Social Normativity

Note first, it is highly likely that language in general and the capacity to learn and apply norms co-evolved:

> "[N]ormativity and language, two hallmarks of human cognition, are intimately related. Language acquisition, in particular, is greatly assisted by norm-governed social institutions. These institutions probably had an even greater role than they

> have today early on in the evolution of language. Human language and normativity transformed and shaped each other during their evolution. They remain closely intertwined in contemporary humans."
>
> (Lamm 2014: 283)

The overall story here is that there is a whole range of capacities that humans must have either inherited from their ancestors or gained at an early stage of their evolution allowing them to sustain the complex social world they share. This is often exemplified in the literature by reference to all the things young children have to learn in order to become socially functional adults. To exemplify, imagine a one year-old girl in the fruits section of a supermarket, staring at an enormous pile of what apparently are delicious apples right in front of her. She also sees how people, including her father, pick up several of these apples and claim them for themselves. Then she does what must seem to her like the only sensible thing to do: grabbing an apple and happily sinking her teeth into it. But much to her dissatisfaction, she is quickly interrupted by her father, who says something beyond her comprehension about first having to pay for things in a supermarket. What this little story illustrates is that, in order for our social lives to function properly, we need a shared understanding of the social world we have created on top of the physical one. That is, in order for the supermarket to exist and function as a supermarket, people need a shared understanding of what a supermarket is and how it works. Likewise, they need a shared understanding of their traffic system in order to even get to the supermarket in one piece. An important component of this shared social world, this "common knowledge", are norms. (Chwe 2001: 26) In the case of the supermarket, for example, there is a rule permitting customers to collect all of the things that are offered on the shelves. There is also a rule requiring all customers to pay for the things they have collected with money before they take them home and consume them. There is further a rule clarifying that the exact amount of money customers have to pay for each thing is non-negotiable and determined by the supermarket.

What we learn from this is that the norms we live by are not isolated rules in a rule book of social life for some society X. Rather they are embedded in a rich web of social stories and knowledge. Cristina Bicchieri points to this fact when she says that social norms are triggered by certain cues in our environment (such as the sign of the supermarket outside) which in turn trigger the mental supermarket schema and script of appropriate behavior (such as *Get a cart!*, *Select some things!*, *Pay for them!*, *Say thank you and goodbye!*, *Pack everything onto your cargo bike!*, *Return the cart!*, *Ride home!*). Social norms are embedded in such scripts, which in turn are embedded in the general and shared interpretation – the mental schema – of a given scenario. (Bicchieri 2006: 81-99)

Young children have to learn about the social world and its rules, just as they have to learn about the physical world and its rules in order to become fairly autonomous individuals. Likewise, as they learn that a hammer is a tool for manipulating the physical world, they learn by means of "pretend play" that language is a performative tool to manipulate the social world. (Wyman 2014) So they may bring you a toy and tell you *It's a gift!* and then expect you to be happily surprised, say *Thank you!* and enjoy your new belonging. They do not really mean that they want to gift you their toy. Quite to the contrary, they will most definitely want their toy back eventually. They are rather practicing how declaring something to be "a gift" actually changes something about the thing that is gifted in the social world and how this has further social implications, although nothing physically about the gifted object is changed at all. It is hard to imagine how such a complex, shared (normative) social world could exist without the kind of complex language we have. Also, looking at things the other way around, there would be no need for complex language without a complex, shared social world. Thus the strong presumption in favor of a co-evolution of language and norms.

Reason May Not Matter

But what about reasoning more specifically? Essentially, the whole tradition of public reason and public justification rests on the assumption that reasons and reasoning are important for social order. But there are at least two reasons for being skeptical that they are.

One reason for being skeptical about the project of public justification is that efficient social order, understood as internalized norms, is not necessarily dependent on individuals having good reasons. To see this, note that all the abstract models of how norms solve a problem of cooperation and coordination do not require any kind of conscious reasoning on part of the individual norm-follower. All that is required is that individuals are equipped with a norm system that automatically takes care of norm acquisition and norm implementation:

> "The function of the acquisition mechanism is to identify behavioral cues indicating that a norm prevails in the local cultural environment, to infer the content of that norm, and to pass information about the content of the norm on to the implementation system, where it is stored and used. [...] The implementation mechanism performs a suite of functions, including maintaining a database of normative rules acquired by the acquisition mechanism, generating intrinsic motivation to comply with those rule as ultimate ends, detecting violations of the rules, and generating intrinsic motivation to punish rule violators."
>
> (Sripada and Stich 2006: 288-289)

To make our image of simple norm-learners and norm-followers more human-like, we can also assume that they do argue with each other extensively, but that this does not really influence the norm system, because the norm system only responds to observed behavior. Note also that our little tale of unconscious rule-following must be at least part of what is actually happening, because we do learn and apply norms subconsciously. Without a doubt we are socialized into a world of norms that we have not actually chosen and we continue to follow them independently of any reasonable reflection. At least until they are brought to our attention by some disruptive event.

Another reason for being skeptical about the importance of well- reasoned justification of norms is our tendency to produce bad, self-serving justifications. Clever experiments have often exposed our reasoning as a post-hoc fabrication of why we did something that does not really add up with the facts. (Haidt 2001) So generally, when reasoning by ourselves we tend to fabricate a story that fits well with the image we like to have of ourselves or that we think will appeal to others. But this story often does not track the actual reasons for why we do or believe something. Another aspect of this problem is that we are heavily biased toward things that are in accordance with what we already believe in or value. Optimistically, people reason like clever attorneys but not like judges. (Rosenberg 2014: 104; Uhlmann et al. 2009) Pessimistically, "human reason is both biased and lazy. Biased because it overwhelmingly finds justifications and arguments that support the reasoner's point of view, lazy because reason makes little effort to assess the quality of the justifications and arguments it produces." (Mercier and Sperber 2017: 9)

Taken together, the possibility and reality of a subconscious norm system and the flaws in our capacity to reason pose a considerable challenge to any account of public justification. At least they challenge us to explain why and when reasoning and justifying are more than a rationalization of the given norms we have to live by.

Why and When Reasons Matter

Although there is some truth to the skeptical points just mentioned, the devastating conclusion that reasons do not matter does not follow. What they do show is that overconfident rationalist perspectives, presuming that everything decisive in social normativity has to do with reasons, is clearly mistaken. Rather, what is needed is a more complex model integrating reasoning, emotions, subconscious processes and socialization. (Haidt 2001: 828)

The core problem of this task is the integration of two distinct perspectives. The first one is the socio-biological perspective of the rule follower, born into and shaped by a set of preexisting norms. The second one is the internal perspective of a person, reasoning about and choosing norms for herself.

In his forthcoming contribution, *Two Ways to Adopt a Norm* Daniel Kelly [7] turns to this problem by distinguishing between *avowed norms* and *internalized norms*. Avowed norms are norms that we think of as self-chosen norms such as the norm *Don't eat meat!* after having come to the conclusion that factory farming is wrong, and thus one should be a vegetarian. Internalized norms on the other hand are a special kind of non-avowed norm. "They are socially acquired behavioral rules stabilized by communal practices of intrinsically motivated compliance and enforcement." (Kelly n.d.: 3) Internalized norms are of particulate interest because they come with their own motivational resources and appear to be stored in their proper psychological system. They point to a powerful innate system for human coordination and cooperation ready to be filled up with norms.

Kelly further points to the importance of linking our phenomenology of norm following and the distinction between avowed and internalized norms with cognitive science. Here, a typical process-oriented differentiation is the distinction between "System 1" processes that are automated, fast, intuitive, and effortless, and "System 2" processes that are slow, deliberate and guided by effort and attention. (Kahneman 2011) Now, internalized norms in the norm system are clearly an instance of System 1. But our personal experience of reflecting, deliberating and perhaps somehow even choosing avowed norms clearly extends to System 2. The main challenge, then, seems to be coming up with an account that explains norm following in terms of both systems and their interaction.

There are accounts that try to live up to the challenge. Kelly points to Victoria McGeer and Philip Pettit (2002), who argue that what distinguishes humans from other minded beings is their capacity for self-regulation. This capacity allows us to self-select constraints in contrast to simpler, merely "routinized minds", that are operating only under exogenously given constraints. Unsurprisingly, the main vehicle of self-regulation is language, as it allows us to mentally attend the content of constraints and their implementation. However, the authors are primarily concerned with epistemic questions and not with whether and how the self-regulated mind can change the norms that constrain the routinized mind. Thus, on the account of McGeer and Pettit (2002), the question remains as to how expressing an avowed norm such as *Don't eat meat!* can lead to a state where the person adopting it is internally motivated to follow it, as would be the case if it somehow were internalized.

As far as I am aware, to this day there is no systematic psychological account of internalizing avowed norms. There is of course plenty of anecdotal evidence that something like this must be possible. Think for instance of the person that becomes a vegetarian at some point in life, changes her behavior and develops a deeply-rooted aversion to eating meet in general, to

[7] This forthcoming paper is available at Daniel Kelly's website: https://web.ics.purdue.edu/~drkelly/. Page numbers refer to this unpublished version of the paper.

the extent that such a person might feel disgust and anger at the sight of others feasting on pork ribs. Also experiments have repeatedly shown that "norm talk" does make a difference. (Bicchieri 2006: 153 ff. Shank et al. 2019) What is missing is a deeper understanding of how we can reason ourselves to new norms. More precisely, we do not know how and whether avowed norms can become internalized norms. Thus we have to leave the matter to further (psychological) research.

From a more sociological perspective, however, there are some insights into how effective norm change is possible. One such finding is that pointing to internal inconsistencies in our norms, behaviours and the reasons we enlist in their defence, does make a difference. (Mercier and Sperber 2017; Bicchieri 2017; Summers 2017) Accordingly, invoking consistency arguments at least allows us to change how internalized norms are applied. For example, if you have the internalized norm to prevent suffering, and I convince you that animals suffer just as humans, this may lead you to treat animals differently in order to remain consistent. Another interpretation of this example would be that preventing suffering is a deeply held value, rather than a behavioral norm, and that we seek behavioral norms that are consistent with our values. Another common finding is that, given certain favorable conditions, group discussion can be an effective way for changing norms. Hugo Mercier and Dan Sperber (2017) argue that our capacity to reason has evolved in order to justify ourselves and to evaluate the justifications of others. Thus, in the right social setting (preferably in small groups), reasoning can allegedly overcome the flaws it exhibits when practiced in solitude and can fundamentally change how people think and act. To support their case, Mercier and Sperber (2017) point to historical evidence of changing norms, such as the abolition of slave ownership and experiments in political deliberation by James Fishkin.[8] Pointing to insights from interventions in small communities (e.g. with the goal of changing gender norms and practices such as female genital cutting), Bicchieri (2017) also argues that group discussions are a powerful tool for changing norms. In light of her theory of social norms she further argues that the core mechanism driving norm change is changing social expectations:

> "Group discussion has an important public dimension. During these discussions, people's acceptance of certain arguments becomes visible, which may induce participants to be more willing to accept such arguments themselves. Discussion helps to change our personal normative and factual beliefs and to observe that others'

[8]Their reference is to James S. Fishkin (2009): *When the people speak: Deliberative democracy and public consultation*, Oxford.

> beliefs are changing, too. The process of belief change becomes a collective one, as we change our minds together."
>
> (Bicchieri 2017: 161)

Essentially, Bicchieri holds that under favorable conditions (group diversity, equal rights to speak, no power asymmetries, no violation of taboos or deeply held values), publicly abandoning some norm or agreeing on a new norm and on respective sanctions, changes expectations and beliefs. The important bit here is that the publicness of the activity creates higher order expectations and beliefs. As an example imagine that a group of individuals agrees to change their greeting norm from shaking hands to bumping elbows, because this reduces the risk of spreading a dangerous disease. Assuming that all group members come together and verbally confirm this and also agree on a punishment for deviations from the new greeting norm, this creates are range of higher order beliefs and expectations: After the discussion, every individual believes that everybody has good reason to follow the new norm. She also believes that the others believe that she herself has good reasons to follow the norm. Knowing this, she may form the expectation that others will and ought to follow the norm, just as she expects others to expect that she will and ought to follow the norm.
From the work of David Lewis onward it has been a common theme in the literature that such higher order expectation, facilitated by common knowledge, enable stable coordination. Group discussion seems an effective way to establish and manipulate social expectations and shared beliefs. Bicchieri further suggests that if we succeed in collectively adopting a new social norm, compliance may eventually becomes habitual and indeed, internalized. (Bicchieri 2017: 117-118)

What do we learn from these reflections on the relation between reasoning and norms? Firstly, it does seem likely that there is a System 1 type norm system that enables us to pick up on observed norms and internalize them. It is unclear whether a similar system exists as a System 2 process guided by reasoning. Assuming that two norms systems actually do exist, we still do not know how they interact. Secondly, irrespective of this gap in research and the mentioned limits of human reason, reasoning and in particular reasoning well in groups does make a difference for the norms we live by. More specifically, public discussions are one way of collectively changing shared beliefs and explications and thus changing social norms.
Overall, this suggests an indirect connection between reasoning and internalized norms: Our norm systems do not respond to reasoning directly, but to what we observe to be the common behavior in our community. What we can do, however, is to collectively change the social norms that shape actual behaviour. Then the norm system might pick up on this change and fully

Figure 4.1: A Simple Two-Stage Game

Stage 1: A Cooperation Problem

	Cooperate	Defect
Cooperate	$2, 2$	$0, 3$
Defect	$3, 0$	$1, 1$

Stage 2: A Coordination Problem

	Norm X	Norm Y
Norm X	$3, 2$	$1, 1$
Norm Y	$1, 1$	$2, 3$

internalize the new norm. In a nutshell, persistent coherence in reasoning *and* action might be decisive: Individuals internalize a social norm if their community manages to collectively reason and act upon it.

4.1.4 Concluding the Empirical Model

This also concludes our inquiry into the evolution of cooperative social order. What we have seen are the contours of a theory of social order that emerges out of a wide range of existing literature in biology, philosophy, psychology, behavioral economics and related fields. So far, this "theory" is no more than an accumulation of related abstract models and piecemeal experimental findings. However, I have depicted only the tip of the iceberg of this highly dynamic and productive field of research.
Now, before we turn to the next section and lose ourselves in normative modeling, I summarize the core lessons to be drawn from this section in view of constructing a normative theory of justified social order.

1) What Norms Do for Us – A Simple Two-Stage Model: In *Subsection 4.1.1* and *4.1.2* I have discussed problems of cooperation and coordination. In both cases, social norms have resulted as a key answer to how these kinds of problems are solved in society. Hence, the most general conclusion of this section is the image of social norms as simultaneously solving a problem of cooperation and coordination. You can think of this in terms of a simple two-stage game consisting of a prisoner's dilemma and an impure coordination game depicted in *Figure 4.1*.[9]
If the first game is solved by establishing a cooperative equilibrium, we – so to speak – zoom into the top left quadrant of the cooperation problem - the cooperative solution - and discover that

[9]Bicchieri presents a similar notion by saying that a prisoner's dilemma is converted into a coordination game. (Bicchieri 2006: 26 ff.) Jeremy Waldron also uses a similar illustration. (Waldron 1999: 101 ff.)

there remains a problem of coordination: choosing which specific cooperative norm we want to have in place. This two-stage game summarizes the general function of social order understood as a set of social norms: Providing a stable order that coordinates people's expectations and actions. Assuming that this is a close-enough model of the core function of norms of social order, we can understand why an elaborate apparatus for learning and stabilizing norms could be the result of an evolutionary dynamic.

2) A Model Individual: What does the typical human being, sustaining the normative practices talked about in this section, look like? Well, in respect to her capacity of practical reason, she is several things at the same time. First of all, she is a norm follower. That is to say that she has a norm system which allows her to learn, internalize and implement norms. The implementation relies on a system of emotional cues that also provide internal motivation to follow a norm and sometimes even to sanction observed misconduct. Although our model individual has learned many norms and the extensive social knowledge that is associated with them through socialization, she is not merely a blind follower of the existing conventions. As soon as her capabilities of using language and reason are sufficiently developed, she can, secondly, become a partly self-regulating being, reflecting upon and choosing her own norms – at least so far as society and her own psychology allow for it. Her capability to reason is best applied to reasoning with others, which is convenient because it is difficult to change social norms all alone anyway. Besides being a partly self-regulating norm follower, our model individual is, thirdly, instrumentally rational in the sense that she seeks to maximize the satisfaction of whatever ends she sets for herself. Experimental evidence suggests that she is especially likely to fall back onto instrumental rationality when she believes that there are no norms at work in a given situation. (Hoffman et al. 1997) In summary, we may say that our model individual has a "modular" capacity of practical reason. It is "modular" in the sense of being made up of several modules: rule-following, self-reflection and instrumental rationality. Given that the practical reasoning of our model individual is at least influenced by these three modules, it is difficult to come up with a unifying model that can predict her behavior, not least because the judgments of one module might not always be conclusive and because different modules may demand different behaviors.[10] Apart from the number and content of the modules, the many-modules structure itself does seem like a plausible and perhaps lasting description of a mind that has been and continues to be gradually changed by evolution.

[10]The image of modular reasoning I am painting here is not to be confused with modular rationality in game theory but rather refers to biological accounts of a modular (human) mind as proposed by Peter Carruthers (2006: chap. 3).

3) A Set of Feasible Norms: The norms we live by, according to the literature cited in this section, are nothing more than conventional equilibria even if we, from our own perspective as norm followers, might feel strongly about them. Feeling strongly about norms is simply a mechanism of internal norm enforcement that we experience. It is not any kind of evidence that some norms are special or better than others from some objective, trans-social perspective. Different societies have and do live by very different norms. And although there are types of norms that reappear in all stable human societies (norms that prohibit killing, physical assault and incest, as well as norms of fairness and assistance), the actual norms specifying these broad categories are quite diverse. (Sripada and Stich 2006: 281-282) The scientific theory of social order will probably never explain, let alone predict, norm choice of a particular society. This is because actual norm selection is driven by the particular culture, history and power relations of the society in question. What this leaves us with is the hopeful image that for most human communities with some given distribution of individual values and preferences, there also exists some set of alternative cooperative norms that they can, in principle, coordinate on.

4) The Lost Ideal of Mutual Benefit: The simple two-stage game above suggests that whatever norm is selected in the coordination game, everybody will benefit from coordination. This in turn might lead one to conclude that social norms must always be a matter of mutual benefit. But what does that actually mean and is it true? As long as we are only considering the emergence of highly cooperative social order and norm following from the perspective of biological evolution, the answer is clear: Mutual benefit simply means that the average reproductive fitness of each individual is likely to be higher in a community of sophisticated, conditional cooperators. So in the context of biological evolution, there does seem to be specific mechanism, i.e. maximization of reproductive fitness, which allows us to explain why stable cooperative norms might have evolved. At least this is what the replicator dynamics of Brian Skyrms (1996) or Ken Binmore's (2005: Chap. 7) theory of kin selection are telling us. Correspondingly, we might say that under favorable conditions, everybody can gain from cooperative norms in terms of reproductive fitness: every individual or, more precisely, every gene, gets a higher chance of reproduction in a community with stable cooperative norms. This relatively clear image dissolves if we also consider cultural evolution. Large-scale, flexible cooperation requires a whole range of advanced cognitive capabilities. And as it turns out, these capacities allow us to be, to some extent, self-regulating. As such beings, we may replace concerns for reproductive fitness with, say, an overall concern for a steady growth in GDP or, strange enough, a preference for small families. (Ihara and Feldman 2004) What this means is that in the context of cultural evolution, there is not one specific underlying mechanism driving norm selection that we know of. Hence, neither is there a guarantee that the norms in

place are, in some meaningful way, to our benefit. Rather, they may turn out to be not only public goods but also public bads – i.e. things that are detested by some or even all, but are nevertheless upheld by symmetric, self-enforcing expectations.

5) The Perspective of Justification: Besides these four insights from the emerging theory of social order, I made three key empirical assumptions in *Chapter 1*, which frame the entire inquiry. I restate these assumptions here in order to have a complete and concise compilation of the empirical model, informing the normative theorizing that is to follow.
The first assumption is that human beings generally live under some instance of social order and that the advantage of having such an order is so obvious that it does not require justification. That is, while many individuals may detest some, most or even all norms they live by, it is extremely rare that an individual prefers to live alone in the woods instead of excepting some set of common norms which allow her to live in community with others. Thus, the relevant normative question is not whether social order is justifiable per se, but what set of norms is justifiable to what set of individuals.
The second assumption is that model individuals endorse a requirement of public justification. That is to say that they want to live by a set of norms that everybody can endorse for their own good reasons. This scenario is desirable to our model individuals, because publicly justified social order is likely to be a stable and efficient order they value.
The third assumption is that the set of individuals we are dealing with is *diverse*. This means that they hold different views of the world, the good life and the good society. Consequently, they are divided by deep and reasonable disagreement on many things and especially on the norms that should make up their social order. Of course, this is not the correct empirical description of every human community. But it is the correct description of those communities that concern us in constructing a general theory of justified social order. What I have in mind here are large, anonymous societies, inhabited by individuals with different social and cultural backgrounds. However, many of the insights and arguments presented in this chapter also apply to less diverse communities. The main difference would be that with less diversity the normative perspective constructed would be less abstract and open-ended.

4.2 The Normative Model

As I take it, normative theorizing is always an integration of an empirical model, i.e. an understanding of the actual social world, and an ideal, providing a normative perspective on the empirical model. As the ideal usually also takes the form of an abstract image of how society, or some aspect of it, should be organized, we can also speak of the ideal as a model. Accordingly, normative theorizing is the integration of an empirical and a normative model.

"Integration" means that there is an interactive dynamic. In one direction, we are looking for a suitable ideal from the perspective of the empirical model, i.e. from an understanding of the actual social world, its inhabitants and their problems. In a way, the ideal we then construct must already be part of the empirical model, or at least be implied by existing normative thought. Where else would it come from? Once settled on an ideal, we can then look into the other direction, back onto our empirical model and explicate the ideal in terms of what it implies for our social world. Perhaps we do not only go through this motion once, but rather engage in a back and forth thinking between the two models for some time. Hopefully, what we end up with is a coherent theory of practical relevance because it has devised and never lost sight of an accurate empirical model.

Now that we have a fairly comprehensive understanding of the core empirical model and our object of justification – i.e. social order understood as a set of social norms – it is time to turn to the detailed elaboration of a suitable normative model – an ideal. As you may recall from the last chapter, my image for the core normative model is Political Equilibrium: An open-ended, procedural ideal that assumes justified social order to be some kind of compromise. As you might also recall, many important details of this model are still to be filled in. Also, since this and the following chapter are meant to exemplify how Embedded Constructivism could work, it remains to be shown that the empirical model and the normative model fit together in a coherent and fruitful manner. So the two tasks of this section are, one, taking the first step in the integrative process by moving from the empirical to the normative model, and two, specifying the open-ended ideal of Political Equilibrium.

4.2.1 From Description to Prescription

How does one get from a prescriptive to a descriptive account? One elegant way of doing this consists in identifying a normative problem that is already present in the empirical model. Let's do precisely that by looking again at the image the empirical model has left us with: Human beings are as much socially oriented norm followers as they are rational maximizers of their own expected utility. Human communities are thus usually able to coordinate their actions according to some set of cooperative norms. They can further (collectively) reflect upon and reason about which set of norms they want to live by. That is, they can engage in what one may call politics, broadly conceived. But as they do this, they face the most basic version of the problem we are concerned with in this inquiry: What norms *should* they choose? So there it is. Our little story of human beings as norm followers has, quite naturally, led us to a normative problem. I say "naturally" because as partly self-regulating norm-followers, we have the natural capacity to engage in normative reflection about our social world. At the same time, the cognitive capacities that are necessary for being partially self-regulated seem

to have unleashed us from the original mechanism that once guided their creation and our behavior: mutual gains in terms of reproductive fitness. This unleashing pretty much forces the self-regulating being to reflect upon and choose the norms it wants to live by. It is like considering whether to stay at home or go outside for a walk: Once you have started thinking about it, you cannot not make a choice. Likewise, as we grow up and realize that the rules that surround us are a human creation and could be different, it is almost impossible to not form a political stance toward them, even if that stance implies staying at home on election day.

So as human beings we are, quite naturally, facing the individual normative problem of evaluating the norms there are and the norms there could be. This translates into the collective basic problem of politics: What norms should we live by? Which, more theoretically speaking, translates into the problem of: What norms should we select from the eligible set of norms?

Reintroducing Justification

Having specified a normative problem, we further need a normative perspective in order to get out normative model off the ground. The normative perspective we turn to here is that of justified social order as laid out in *Chapter 1.* There I argued justified social order is an highly desireable ideal for our model individual. As a guiding principles for explicating this ideal I have further put forward the justification principle for social norms (JPN):

> **JPN**: A social norm N is justified to an individual i in society S governed by that norm to the extent that N being a positive norm in S is coherent with preferences of i, given that
>
> 1) i has formed well-considered preferences on social order,
> 2) N being a positive norm in S is strictly preferred by i to having no social norm governing the domain of N in S,
> 3) i is at liberty to openly reject the JPN in S.

By itself this principle is not of much help to our model individuals, having to choose the right norms to live by. The hopeful message of our reflections on Political Equilibrium in *Section 3.2* was that there likely exists a set of eligible justified norms for most societies. In more detail, the claim defended by Gerald Gaus in *The Order of Public Reason*, is that while there is no agreement on the most preferred norm(s), there are cooperative norms that are, if followed, uniquely justified to everyone. The task now is to explain why this holds for the model individuals depicted in the empirical model.

Let us begin with the assumption that our model individual, let's call her *Anna*, does take the time to reflect upon and reason about the norms she wants to live by and actually comes up with a preference ordering. If we now further suppose that Anna is like us, born into an existing social order, inhabited by a diverse bunch of people, she will have a familiar experience: Her most preferred social order, her "utopia", does not match with existing norms and the norms other people prefer – at least not perfectly. Upon experiencing this, Anna might react in different ways. She might revolt, despair, or run off in search for her perfect utopia elsewhere. If Anna has a more pragmatic inclination, however, she might also wonder what kind of ideal she can realistically hope for. What is the next best thing if she cannot get her utopia?

The answer to our model individual's question is what really gets us into the kind of account I defend in this section. It is based on the perspective of the two-stage model of social norms and Gaus' hopeful message of the existence of an eligible set of justifiable norms. Simply put, the claim is that it is realistic enough to assume that Anna and others like her can coordinate on some set of cooperative norms that are mutually beneficial, all things considered.

What makes this claim more complicated in detail is that we do not know what counts as "beneficial" for Ana and others like her. In lack of such a standard and a philosophical reluctance to introduce one (such as Rawls' primary goods), actual individual evaluations are the only basis for judging some social state as an instance of mutual benefit.[11] Whereby Condition 2 of the JPN establishes at least a minimum requirement for some norm to be beneficial respective to individual preferences: It has to at least be preferred to having no norm in place at all.

Still, due to the assumption of diversity, Anna and other members of her community will disagree about which norms, satisfying Condition 2, they should select. But my core claim here is that as Anna and others like her notice that in a diverse society people generally do not get their most preferred norms, they will be pragmatic enough to accept *some* set of mutually beneficial norms, all things considered.

Essentially, this is the first leap from the empirical to the normative model. It consists in the claim that all model individuals will endorse some kind of social order if, all things considered and according to their own evaluation, this turns out to also be to their benefit. Accepting social order as a mutually beneficial compromise is, I believe, highly rational and reasonable for a person like Anna, because actual social order is always a kind of compromise anyway. It is a compromise that can yield high gains from cooperation to those who take part in it.

We already discussed the rationality of converging on cooperative norms while disagreeing on

[11] The "philosophical reluctance" to introduce a universal currency of justified social order such as Rawls' primary goods is of course due to the worry that doing so involves the kind of constructions that make HCM so controversial. On this, see my first line of critique against HCM in *Section 2.3*.

the best norms in *Section 3.2*. Particularly, in reference to Gerald Gaus' "Kantian Coordination Game" and my WhatsApp example. The general upshot of the discussion is that coordination in such circumstances is possible and that once a specific norm from the realm of mutual benefit is selected, this norm becomes uniquely justified respective its contenders. This is because established norms, norms that are part of the current status quo, have the advantage of already coordinating actual behavior on mutually beneficial outcomes, whereas changing behavioral patterns is difficult and costly.
Jeremy Waldron makes a similar claim in respect to legal norms. Specifically, he claims that under the "circumstances of politics", i.e. in face of persistent and reasonable disagreement about the correct laws, people should nevertheless be pragmatic enough to respect the law for and when it coordinates us on a cooperative order. (Waldron 1999: 101 ff.)

If I am right about Anna, she will agree that a compromise of mutual benefit is indeed appealing, because there is some conciliation in knowing that the norms she lives by, although probably never being her most preferred norms, can at least be to everybody's advantage and thus a true public good. But upon acknowledging this broad ideal, she will probably point out that this alone does not help her much. She still does not know which exact norms she – together with everybody else – should settle on.

Apology for an Open-ended Ideal
On this we, as theorists, cannot but disappoint Anna. We should acknowledge that the problem of diverse, unknown and probably inconclusive, preferences cannot be solved in theory. Admittedly, I have not proven that this is impossible, but I think my critical arguments in *Section 2.3* have shown that it is very likely that all such efforts will either involve illegitimate abstractions from the preferences there are, or a never-ending philosophical debate about the one correct social point of view from which to decide such matters. Therefore, our disappointing but honest answer as theorists to Anna must be that we simply do not know and probably cannot know what norms she and her fellow citizens should live by.
At this point Anna may demand that we should get together a bunch of social scientists and try to calculate the optimal set of social norms for her, given a comprehensive study of her society's status quo. This, however, would miss the deeper insight from the discussion in *Subsection 3.2.1* of Hayek and the calculation debate. In short, Political Equilibrium, just like the ideal of market equilibrium, is an elusive ideal because it suggests that there is some unique state – one particular equilibrium – that is the optimal social state and the whole idea is to discover that optimal state. But the truly deep point, eventually made by Hayek, is that this state does not exist. There is no unique optimal equilibrium or any equilibrium for that matter, independent

of the social mechanism producing it. This is because first, beliefs, circumstance and thus preferences change. Second, people do not have well-considered preferences independent of engaging in a respective social process. Where else would preferences be formed? What would they be about? How for instance could someone have well-reflected economic preferences without having been to the market?
Essentially, the first two points imply that the eligible set I have been talking about so much does not exist independent of a social process that continuously tries to establish what is in the set. To illustrate this point, think of a big tournament, such as the FIFA world cup. Nowadays, there is a huge amount of data available about the past performance of teams and individual players. So one could, based on this data, build a complex statistical model and simulate the outcome of the tournament. But doing so would completely miss the point of what the tournament is all about. The tournament is about all the things that happen within the tournament. Hence, there is no winner (or loser) of the tournament in any meaningful way before the tournament has actually taken place, including all of its contingencies. A successful tournament does establish a unique solution, but it is *only* unique relative to the particular tournament that produced it. If the tournament were to be repeated two weeks later, a different team might win. And at some point the tournament should indeed be repeated, or the last winner should at least stop referring to itself as the "world champion", because things change. Returning to the world of social norms and the problem of justification, we can conclude that (well-considered) preferences, the eligible set of (justified) norms, and selected norms can only be thought of as dependent variables in a continuous social process. Therefore, the ideal of justified social order must be a procedural ideal aimed at an open-ended social process: "we learn what our ideal is as we seek it." (G. Gaus 2016: 136)

What we can do for Anna is to point to her capacity to jointly reason about the right kind of norms and to the status quo she finds herself in. Nature has equipped Anna and her fellow citizens with the capacity to reason themselves toward justified social order and they do not have to start from nowhere. Rather, if they are fortunate to already inhabit a stable and cooperative order, the norms and institutions constituting their status quo will be their starting point and initial benchmark for reform. What they do not need to do is to seek some philosophically sterile environment. This would probably not lead anywhere. What they do need, however, as they are trying to establish justified social order, is a standard of reasoning well. This is, as we know from Hugo Mercier and Dan Sperber (2017), because individuals such as Anna often reason poorly. And, unsurprisingly, the same is true for groups that do not reason under ideal deliberative conditions. (Bicchieri 2017: 157-159)
This is where the idea of Political Equilibrium comes back in. Essentially, I believe there are two

things we can offer Anna from our theoretical normative perspective in order to get her started on her quest for the right kind of social norms: One, an open-ended and procedural account of justified social order. Two, some insights into how this ideal could actually materialize in her community. However, the actual task of deciding upon a particular political system and selecting particular norms will always remain the task of Anna and others like her. That is indeed a lot to ask from our model individuals. But I believe biological and cultural evolution has equipped them well enough to live up to the challenge. Working through it themselves will hopefully allow them to recognize the norms they live by to be, at least to some extent, *their* norms and not just the dictates of some distant ruler of the political or philosophical type.
In order to get to this hopeful image, however, there still is some work to be done. In particular we are still lacking an account of the range of reasons that are relevant for justification in a theory of Political Equilbirum and the range of justification they allow for. But before we get to this task, a few words of clarification are in order.

An Argument and two Caveats

Our story of Anna and her social order was meant to illustrate that the ideal of justified social order I am developing in this inquiry is well suited and well addressed to our model individual and her problem of selecting the right norms to live by. That is to say, my claim is that the normative model that is beginning to emerge is well suited to the empirical model. However, there are some possible misunderstandings we need to eliminate at this point. In order to get these out of the way, let us first consider the leap from the empirical to the normative model in a concise manner:

1) The Circumstances of Politics: All model individuals stand to gain significantly from cooperative order, while fundamentally disagreeing on the best order.

2) The Fact of Social order: Human beings generally have and continue to live in communities, governed by sets of norms.

3) Therefore, model individuals generally do accept social orders that are mutually beneficial compromises.

One possible misunderstanding is that the argument is meant to show that all people would *always* pragmatically accept social orders as a mutually beneficial compromise. This is likely to be false, but the fact that there are some counter-examples does not justify a rejection of the account.
In more detail, accepting social orders as a mutually beneficial compromise is, I believe, highly rational and reasonable for a person like Anna, because actual social order is always a kind of

compromise anyway – but a compromise that can yield high gains from cooperation to those who take part in it. Nevertheless, a model individual, let's call him Bruno this time to keep scenarios separate, could in principle reject social orders as a mutually beneficial compromise. Perhaps because Bruno is convinced that his most preferred social order is the only acceptable order and any kind of compromise is a scandal. (G. Gaus 2016: 222) Or Bruno might believe that he is so powerful that he can get a better deal in the state of nature. If no argument can persuade Bruno to take the pragmatic step and accept the compromise for the sake of peace, stability and cooperation, we have to let it go. Then, the account of public justification as I am defending in this inquiry is simply ill-suited to Bruno.[12]

I do not deny that there are some Brunos out there in the real world. But I also think that the many stable cooperative orders we have, prove that there are mostly Annas – i.e. individuals willing to accept social orders as a mutually beneficial compromise. And in face of the great amount of different social orders that already have been produced and sustained by human beings on this planet, I think it is safe to assume that human beings are by nature fairly flexible concerning the kind of norms they live by. Also, a single human being is not very powerful irrespective of an existing social order that puts her into a position of power. Therefore, I believe human beings are clearly rather Annas, although unfortunately, some are indoctrinated into being Brunos by religion, totalitarian ideologies, a fear of relativism or the like.

Further, and more generally, I am not claiming that justified social order is necessarily of primary or overriding importance to Anna. Maybe Anna is more concerned with a steady flow of affordable consumption goods than with matters of social order. If she would take the time to think about it, maybe she would endorse some account of justification. But we cannot be sure that she would. Anna might be primarily concerned with fundamental threats to her livelihood or the survival of her community. We also do not know what happens when important goals or values collide with justified social order. Perhaps in some scenario where stable order is already achieved, but achieving a justified order appears difficult and uncertain, Anna would prefer certain stability over uncertain justified social order. Generally speaking, I try to avoid argumentation within people's preference orderings and I have not attempted to show that for people like Anna, justified social order necessarily always ranks higher than other important considerations. Therefore, I am not claiming that justification is necessarily

[12] "Some perspectives are, in the end, unable to share a framework of moral accountability with diverse others. [...] Such "Excluded Perspectives," which cannot find sufficient space in the Open Society, will almost surely be those that are committed to the optimizing stance, or some near approximation to it. Faced with different rules to live by, the Excluded Perspectives can live only by those that they think best, and so they cannot endorse the characteristic institutions of the Open Society, which seek to provide as much space for all as is possible." (G. Gaus 2016: 222)

the first virtue of social institutions.

4.2.2 A Compromise with Ownership

Where are we at this point? So far in this section I have tried to show that the emerging account of justified social order is well suited to the empirical model. Now it is time to unpack the coherence relation between well-reflected individual preferences and justified norms as stated in the JPN. To this end recall that in *Subsection 1.3.2* I have said that a social norm is justified to an individual to the extent that it is *coherent* with her well-considered preferences. This core element has mostly remained a black box until now. But this clearly has to change if we are to fulfill the main goal of this section: clarifying and specifying Political Equilibrium (PE) and thereby our normative model.

The Range of Justification

To get started, first of all we need to bracket the practical matter of how individuals come to have well-considered preferences. This will be one of the things we return to in *Chapter 5*, when we start thinking about PE in more practical terms. For now, let us simply assume that our model individuals have well-considered preferences regarding existing and possible social norms.

Second of all, consider a summary of the necessary assumptions of the analysis:

a) We have a society of model individuals (i.e. partly self-regulating rule followers).

b) All individuals have well-considered preferences regarding existing and possible social orders (Condition 1 of the JPN).

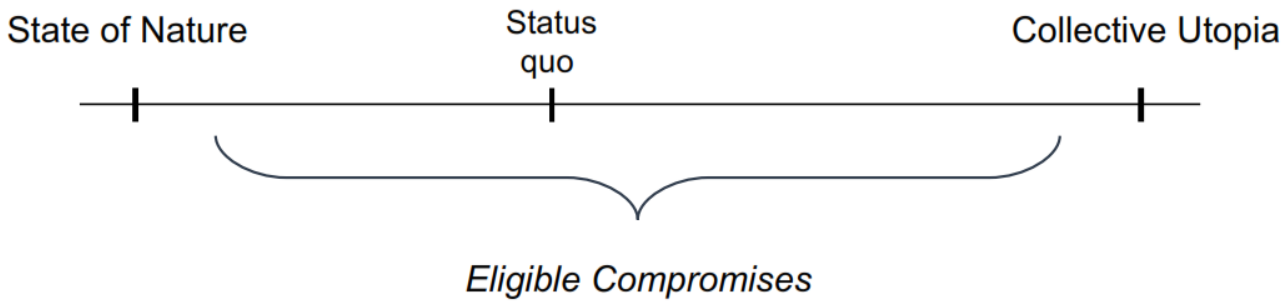
c) Orders that are not preferred to the state of nature by all are to be disregarded as clearly unjustified (Condition 2 of the JPN).

d) Our model society is diverse but not too diverse: Preferences differ significantly, but there is an eligible set of orders preferred by all to the state of nature.

e) Our model society is fortunate enough to have a stable social order.

This gives us a definite minimum point, maximum point and in-between space of justification. The minimum point according to *c)* and *Condition 2* of the JPN is that the social order is marginally preferred by all model individuals to the state of nature. The maximum point is the social order that is the most preferred order of all individuals. However, due to the diversity assumption *d)*, both of these extreme cases and all options close to them are extremely unlikely.

Figure 4.2: Range of Justification I



Rather, the eligible set of social order in light of the assumptions made is likely to be located well between the minimum and the maximum point, defining the range of justified social order. For an illustration consider *Figure 4.2*.

The idea here is to simplify things by thinking of all proposed social orders as being located on a line between the minimum and the maximum point. Thereby we have simply assumed by *d)* and *e)* that our model society is fortunate enough to have a stable order and an eligible set within the range of justification. This is of course not necessarily the case, but I presume that the gains from cooperation are large enough in most cases to justify this assumption.

Now, the difficult bit is to interpret the space in between the minimum and the maximum point. Here it is tempting to interpret the line as a vector of progress: moving to the right means more justification, moving to the left means less justification. Such a vector could for instance be produced by asking all individuals to rank all orders on a scale from 1-10, thus translating their preferences into one comparable scale, and then simply adding up the scores for each order (i.e. applying the method know as *Borda Count*). This would give us a *numerical* social ranking of all proposed social orders for a given society, so that being more to the right on the line and closer to the maximum point means having a higher numerical degree of justifiedness. However, we would of course not really be justified in saying that some order X that is located to the right of some alternative order Y is superior to Y. This is because moving to the right implies that overall justifiedness improves, while it probably also implies that some individuals get a less preferred order. So in order to justify moving to the right, we would further have to introduce a utilitarian principle in favor of maximizing overall justifiedness, rather than maximization of individual satisfaction of preferences. So far I do not see why our model individuals should accept such a principle. More generally speaking, Kenneth Arrow (1950) has famously shown that there is no obvious way of aggregating individual preferences into a

social preference ordering without the violation of some intuitive criteria. Therefore, moving from left to right on the line is to be understood as a mere numerical improvement without normative implications.

The only way to get a more informative ranking in an uncontroversial way would be to restrict the ranking to Pareto improvements respective to the status quo. Then the eligible set would shrink to those, and only those, orders that are located to the right of the status quo and do not make any single individual worse off in terms of preference satisfaction. However, the problem with this model is that it leads to the same kind of problem in the case of having more than one feasible Pareto improvement: There is no obvious way of producing a social ranking of these options. Also, the Pareto criterion is not as uncontroversial as economists are having us believe. To see this, consider that Pareto optimality as a general requirement of such a model rules out the possibility of large improvements in overall preference satisfaction at the expense of making some individuals somewhat worse off. Excluding this possibility altogether would be odd because it is perfectly reasonable for our model individuals to take a large step to the right on the vector of justifiedness while condoning that some individuals are made worse off (although perhaps marginally or only down to a certain threshold). We do this all the time when implementing solidary, re-distributive norms.[13]

In conclusion, the utility principle is not generally justified because it allows for the unreasonable disregard of individual preference, whereas the Pareto criterion is also not generally justified because it does not allow for the reasonable disregard of some individual preferences.

What is certain, however, is that every order within the eligible set is a compromise. It is a compromise because individuals generally do not get their most preferred outcome. This diagnosis is amplified by the consideration that, practically speaking, every complete social order is usually a complex package deal containing different social norms. And each individual may think of the individual members of the overall set more or less highly. Now, this is not to say that individuals may only look at social order as a package deal they can either take or reject. What I do believe is that, if our model individuals are pragmatic enough to accept social orders as a mutually beneficial compromise, they will also accept that social order is always a mixed bag of more or less preferred norms – especially in the case of large societies with complex social orders. Therefore, model individuals accept entire packages of norms if, one, the overall package ranks above the state of nature for each individual all things considered, and if, two, the package deal is not defeated by including outright unacceptable norms.

Condition one implies that single norms may be below the state of nature threshold for some individual because this can be compensated for by other, more preferred norms in the pack-

[13]Pareto optimality as a general requirement is really an anti Robin Hood principle whereby many times, a Robin Hood is really what we need to get to a more efficient distribution in terms of individual utility.

age.[14] This kind of compensation-logic does not apply to all individually undesired norms. Thus, condition two requires that the package does not contain a single rule that is outright unacceptable according to any single individual.

Admittedly, being "outright unacceptable" is a blurry kind of standard. To have an example in mind, consider Rosa Parks' famous refusal to relinquish her seat on a public bus to a white passenger. This refusal exemplifies a rejection of a norm, or rather of a subset of norms, namely all norms facilitating racial segregation and inequality that were part of social order in the USA of 1955. One of the more abstract insights from this example is that social order cannot only be looked at as a package deal. Rosa Parks' refusal was not necessarily a judgement of the overall social order. We do not know if she would have preferred her status quo social order, including racist norms, over the state of nature all things considered and it is really besides the point if she did. The example of Rosa Parks and the Montgomery bus boycott is not about the evaluation of an entire social order, but rather about single norms (and what they represented) that struck some individuals as outright unacceptable to the point that they clearly could not accept them as being part of an overall beneficial compromise.

A typical symptom of such a rejection, as in the case of Rosa Parks, is civil disobedience. Such acts are a deliberate, public violation of a standing norm, whereby the violator anticipates significant personal costs for her transgression, which in turn signals her evaluation of something (e.g. a norm) as being outright unacceptable. So what we generally learn from such cases of civil disobedience is that violations of people's core values[15] cannot be compensated by the benefits provided by other norms or the social order as a whole. Quite to the contrary, outright unacceptable norms will continue to occupy, irritate, outrage and perhaps even alienate single individuals. Therefore, such norms defeat pragmatic overall acceptance of social order as a package deal.

What has been argued in the normative model so far? One, the range of justified social orders is made up of compromises that all modal individuals prefer to the state of nature while not containing any defeaters – i.e. norms that violate core personal values or needs. Two, the social choice perspective of aggregating individual preferences is very helpful in understanding the basic problem of justifying social order, however, it is not helpful in providing a solution. As we have seen above, even if we only look at Pareto improvements, this does not get us very far.

[14] For example, consider that our legal systems have become so refined that we could easily find laws in a domain where we personally find it completely unnecessary to have a regulation. But such "unnecessary" norms do not challenge the overall beneficial nature of the legal system at all.

[15] These are more stable and profound than other personal values and goals. (Bicchieri 2017: 159-161; Schwartz, Caprara, and Vecchione 2010)

Being More Optimistic

The range of justification we have discussed so far is still very broad. Essentially, we have only defined an unrealistic utopia of everybody agreeing on the best social order and a pessimistic minimum that is a kind of pure modus vivendi – i.e. everybody marginally prefers social order to not having any order in place. I have further said that for our model society the eligible set of justifiable orders probably lies somewhere well between the minimum and the maximum point. But although I believe this to be a quite realistic depiction of what they can expect, at the same time it is not saying much. In particular, it is not providing an ideal in the classical sense: a positive goal to work toward. Fortunately, there is nothing keeping us from being more optimistic *within* the realistic range of justification.

In order to do this reasonably, let us distinguish two kinds of reasons our model individuals have, endorsing social order as an overall beneficial compromise: *pragmatic reasons* and *personal reasons*. Personal reasons favor something in virtue of corresponding to what you personally want for yourself. Hence, your personal reasons point you to what you consider to be of intrinsic value. For instance, as a bicycle enthusiast, you might be looking for an awesome road bike. The awesome road bike you are seeking features, needless to say, a classic titanium frame, high-quality components and is offered at a reasonable price point. Now, further assume that you happened to stumble upon a bicycle much like this at a yard sale. This bike meets all of your desires, except that some of its components are not black but silver, which, of course, looks a bit pretentious on a titanium frame. Still, in this case you have many personal reasons speaking in favor of buying the bicycle: It has a classic titanium frame, high-quality components and a price you are willing to pay. Yet there is one aspect to the bicycle, namely its silver colored components, that does not correspond with what you personally would like to have. This is where the *pragmatic reasons* come in. In our example, where you have significant personal reasons in favor of buying the road bike in question, you might have a pragmatic reason for accepting the silver colored components as well, given that being silver colored is not a defeater for you. That is, if silver colored components on a titanium road bike were outright unacceptable to you and you would not have any money left to exchange them for black ones any time soon, this might render the whole deal ineligible. If, however, having some silver colored components is not a defeater in this way, you had undefeated pragmatic reasons to *accept* them as a means for satisfying your personal reasons. Pragmatic reasons favor something not because that something is what you want for yourself, but because that thing is a means for getting something else that you personally want. In other words, pragmatic reasons point you to things you value instrumentally, i.e. as an instrument for getting what you value intrinsically. Depending on your balance of pragmatic and personal reason, you then may or may not have sufficient reasons – a well-considered preference – for buying the road

bike.

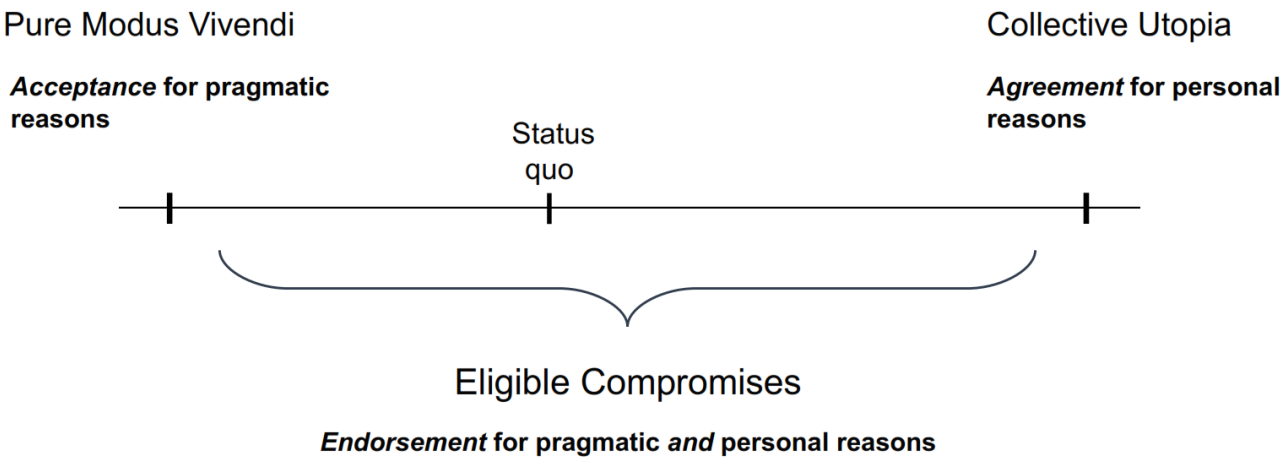
With this distinction in mind, we can deepen our understanding of 'compromise' and restructure the range of justification in a qualitative rather than a numerical way. A compromise, then, is an arrangement where the parties involved are confronted with an option that is not their most preferred option, but which does provide them with a mixture of pragmatic and personal reasons in its favor. This is also to say that an option that is only favored in virtue of pragmatic reasons is, on my account, not a compromise. Think for example of two warlords who have been at war with each other for a long time and both desire, more than anything else in the world, to destroy the other. Nevertheless, as it happens, their respective resources for maintaining the conflict have been completely depleted and they are both running the risk of being overtaken by other, less powerful warlords in the region. So they both agree to a ceasefire. Clearly, their ceasefire agreement is a pure modus vivendi agreement, sustained for as long as their pragmatic rationality converges on this point. There is nothing in this agreement that the two warlords want for themselves. What they really want is to destroy the other and as soon as they will see an opportunity to attack, they will take it. So a pure modus vivendi agreement does not offer the parties anything they want for themselves, which is why it is inherently unstable.[16] I think both of these features are in contradiction to what we take to be a compromise in politics: A resolution where the parties involved have to move from their most preferred option but still achieve a stable outcome because everybody gets at least parts of what they wanted for themselves. Informally put, a compromise is an arrangement where you do not get exactly what you want, but you do at least get parts of what you really want. The other parts you accept as means for getting what you want.

So now that we have a refined understanding of 'compromise', we can also reframe the range of justification in terms of the distinction between pragmatic and personal reasons. Consider *Figure 4.3*.

The core idea of this reframing consists in adding a qualitative specification to the coherence relation between norms and preferences. This specification consists in emphasizing the role of personal reasons for justification. Accordingly, the minimum point is now a pure modus vivendi arrangement that is only preferred to having no social order in place due to pragmatic reasons. Strictly speaking, this arrangement is *minimally* justified because individuals – such as our two warlords – do have reasons to favor it. Thus there is minimal coherence between individual preferences and norms. Nevertheless, this arrangement is highly sub-optimal from an individual and a societal perspective because no one really wants it as such. It is merely the outcome of very peculiar circumstances. And it is very unstable because individuals will not be inclined to sustain it for any longer than absolutely necessary given the circumstances.

[16] A less dramatic example is provided by John Rawls (1993: 147).

Figure 4.3: Range of Justification II



On the other side of the divide we have the maximum point where social order is perfectly justified because the coherence is maximized in a way that personal reasons for certain norms and the norms there overlap perfectly: Every individual gets her most preferred norm, which is only possible if the most preferred norm is the same for everyone. And again, since both end-point arrangements are quite unrealistic – one being too pessimistic while the other is too optimistic in light of our model individuals and their society – we should expect that the set of eligible compromises lies well between them.

What is new about reframing the range of justification in this way? Well, the peculiar feature of this perspective is that it allows us to see justified social order as set of different coherence relations between the order in question and the balance of personal and pragmatic reasons held by different individuals. The intuitive idea, doing the normative work, is of course that as the fit between individual personal reasons (individuals goals, values, needs and norms) increases, so does the individual ranking of theses norms. Thus a compromise gets better in the eyes of each party as individuals get more of what they really want and less of what they merely accept in order to get it.
As a consequence of this perspective on diverse society, the coherence relation between individual preferences and their social order will also be of a diverse nature. In theories of hypothetical choice modeling and in particular in Rawls' theory we are presented the image of social order that is justified in the same way to all citizens. That is, in the Rawlsian paradigm all citizens endorse the order of well-ordered society for the same reason; they share a conception of jus-

tice. From our perspective of social order as a compromise accepted or endorsed for a mix of pragmatic and personal reasons, things are much more diverse. That is, from this perspective, we would expect that the content and mixture of reasons that relate individuals and social order differ. This is of course not to say that there are no cultural patterns here. Societies are prone to have certain narratives of the virtues of their order that tend to be socialized into its members. Nevertheless, by definition, members of the diverse society are not culturally homogeneous and thus we will continue with the image of social order as being justified in different ways respective to different individuals.

What has been argued in the normative model so far? One, the social choice perspective of aggregating individual preferences is very helpful in understanding the basic problem of justifying social order. However, it is not helpful in providing a solution. Even if we only look at Pareto improvements, this turns out to be an unjustifiable straitjacket. Two, we begin to get a more differentiated understanding of the range of justification if we distinguish between pragmatic reasons for accepting and personal reasons for endorsing social order. Thus far the claim is that justified social order consists in a compromise that all prefer to the state of nature for some mixture of pragmatic and personal reasons. Thereby, crucially, that compromise may not contain any defeaters – i.e. norms that violate core personal values or needs and thus defeat the pragmatic logic of the compromise. Nevertheless, the specification of pragmatic and personal reasons itself does not get us much further in that it does not shrink the eligible set of justifiable social orders. Essentially we are still only saying that some order is more justified if an individual values it higher than some alternative. But in specifying what it generally is that makes some compromise better, we can take a further step toward a more optimistic, but realistic ideal.

In order to get there, recall the issue of some norm being a defeater. This is to say that some norm is outright unacceptable because it violates fundamental personal values of some individual. This actually happens. What also happens is the opposite phenomenon that a norm of social order directly embodies what is valuable to us as citizens. This is not just a coincidence, but a core promise of democracy. Such things as basic freedoms and rights, political participation and the rule of law are not just means to an end, but things of inherent value to people who see themselves as democratic citizens (as "We the people"). Therefore, I believe a fitting and optimistic view of justified social order consists in an arrangement where individuals recognize social order as inherently valuable, in spite of it being a compromise. And my hypothesis is that people can actually get there if the compromise is favored sufficiently by their personal reasons.

Now, the first question you might ask at this point is: *How much is sufficient*? The answer

is, I do not know. It probably depends on the person and circumstances in question. We will return to this matter below. A second obvious question points to the fact that the ideal I am suggesting consists in a psychological state and then goes on to ask about the conceptual and empirical nature of this state. This is also a difficult question to answer. One way of answering it would be to suggest that people internalize norms if they consider them sufficiently valuable. This would imply that we can reason ourselves to the internalization of some norm.[17] But the literature relevant to our empirical model was not conclusive on this point. So perhaps our norm system does not allow for a direct internalization by means of reasoning. However, there is another psychological state that is likely to be accessible by reasoning and is conductive to social order: having *ownership* of something.

'Ownership' in Democratic Theory

The concept of ownership has become rather popular in democratic theory. Early on, Philip Pettit argued that the principle of freedom as non-domination requires that "public" decision-making

> "[...] must be a form of decision-making which we can own and identify with: a form of decision-making in which we can see our interests furthered and our ideas respected. Whether the decisions are taken in the legislature, in the administration, or in the courts, they must bear the marks of our ways of caring and our ways of thinking."
>
> (Pettit 1997: 184)

Pettit then goes on to argue that we can "own" public decisions if they are contestable. I share his concern for ownership of citizens for *their* public decisions, whereby I would put the emphasis on the the respective norms, rather than on the decision itself. However, I do not share Pettit's concern for the contestability of public decision and the principles of freedom as non-domination.

What I have in mind is much better expressed by the notion of "democratic ownership" proposed by Cillian McBride (2015). He argues that citizens should have a collective identity as collective political agents and they should have ownership of their institutions and collective decisions. This is meant to ensure that citizens are reconciled instead of being alienated by the fact that collective decision-making often leads to outcomes that are not their most preferred outcomes. Further, McBride argues that public deliberation is the decisive tool for achieving collective identity and ownership of the right kind.

[17] As pointed out in *Subsection 4.1.3*, this seems to be Gerald Gaus' position in OPR.

On a similar note and most recently, Cristina Lafont (2020) argues for an ownership relation between citizens, the state and its laws in light of a participatory ideal of democracy:

> "In what follows, I would like to articulate a participatory interpretation of deliberative democracy that puts the democratic ideal of self-government at its center. In other words, for this conception of democracy it is essential that citizens can identify with the political project in which they collectively participate and endorse it as their own."
>
> (Lafont 2020: 162)

> "They must be able to take ownership over the law and see that it tracks their interests and ideas, their ways of thinking and their ways of caring [...]."
>
> (Lafont 2020: 225)

In accordance with McBride, Lafont sees political alienation as the looming danger for citizens who cannot identify with the laws they live by. She also considers public deliberation as the main antidote to alienation and facilitator of ownership relations.

Leaving specific principles of democracy aside, I believe 'ownership' is just the right kind of conception to think about justified social order, a mutually beneficial compromise, in more optimistic and idealistic terms. More precisely, under the circumstances of politics where we all want to coordinate on some set of norms but cannot agree on the best ones, the best we can hope for is that the compromise that results motivates identification and ownership – perhaps not in respect to every single norm, but at least to social order overall. As I continue to argue in this section, justified social order my achieve this identification in virtue of being sufficiently coherent with people's personal reasons.
This claim, however, as well as the accounts of ownership cited above, leaves a whole range of important (scientific) questions open: What is ownership, psychologically speaking? How does it work in the individual mind and in society? How does having ownership relate to reasoning? Does public deliberation really produce ownership? I cannot answer all of these questions in a satisfying manner. Nevertheless, in the following I attempt to show that we can be a lot more precise about what ownership is and how it works in the context of social norms if we consider the construct of "psychological ownership".

Psychological Ownership

Psychological ownership – from now referred to as PO – denotes an individual, psychological attachment of a person to some object and is discussed at length in organizational science. And

although studies in organizational science are mostly about corporations and the employer-employee relations, this area turns out to be a rich mine of illuminating ideas for our purpose. This is not surprising because organizations are much like a small-scale experimental setup of the broader problem of finding cooperative social norms for large-scale societies. That is, organizational science provides micro and meso insights into the matter of cooperative social order and, as we know from social norms literature, such insights might be just as relevant to the macro level.

The standard reference in organizational science for a comprehensive and pertinent treatment of PO is Jon Pierce and Liro Jussila (2011). According to them, PO is another human capacity that originates with biological evolution but was then taken over by cultural forces in terms of how it is specified and actually applied. The rather obvious origins of this capacity probably lie with behaviors of claiming and defending resources for survival and reproduction: food, shelter, territory, mating partners. But today, human beings can in principle develop PO for all kinds of things – material and immaterial alike. To a large degree, culture and socialization determines what kind of things we end up considering "ours". On a deeper level, PO reflects an identification process between an individual and some object to the extent that the individual considers the object as being part her self-conception. Consequently, she is inherently motivated to defend and improve the object. She also enjoys its possession. In conclusion, PO is a cognitive and affective psychological capacity that we add to the assumptions that make up the empirical model and in particular to the list of things every model individual is equipped with.

Besides PO, there is the construct of *collective psychological ownership*, which is of particular interest in the context of this inquiry because it potentially relates groups of individuals and norms in terms of the ownership relation. Collective psychological ownership – from now on referred to as *CPO* – is psychological ownership of a group of some object X, so that individuals of that group might say that X is "ours" or "belongs to us". Thus, CPO is meant to capture the common phenomenon that groups claim certain lands, values, goals, physical objects or values as "ours" and transform them into a psychological construct that can be studied and explained scientifically. Thereby the standard view is that CPO for X obtains where individuals have, firstly, established a collective identity – a shared mental construct of group membership and group identity, including shared experiences, goals and symbols. Secondly, these individuals think of their group rather than their individual self as the agent of reference when it comes to the ownership of X. Usually, this is established through some collective communicative effort of the group, making or claiming something to be "ours".

Let me remind us of the motive behind all this talk of ownership. So far we only have a very

vague ideal: coherence between social order and personal reasons is better than coherence in terms of pragmatic reasons. But I believe we can be much more precise about why this kind of coherence is a good thing and what it amounts to if we combine it with phenomena of psychological identification and attachment. In a nutshell, my hypothesis is that significant coherence between social order and personal reasons may lead to PO and, adding group identity, to CPO for that order. Combined with the assumptions of well-considered preferences on norms, this provides for a range of highly desirable states and thus for an intriguing and realistic ideal of justified social order. Essentially, turning to PO and CPO allows us to argue for a similar ideal as McBride (2015) and Lafont (2020), but with a much more precise understanding of what we already know and what we don't know. Let us get into the details.

4.2.3 Ownership for Norms

How do PO and CPO connect with the norms we live by? Norms, as almost anything, can be targets of ownership. My core thesis in this subsection is that reasoning ourselves toward having ownership of the norms we live by denotes a suitable realistic and optimistic ideal of justified social order. It is suitable because it is a kind of internalization of norms for the right reasons. It is realistic because we remain in the realm of compromise while knowing that people actually can establish relations of ownership for norms they identify with. And it is optimistic because if we have PO and CPO for the norms we live by, this state brings with it some highly desirable benefits: It allows for an autonomous, active citizenry that is capable of advanced forms of cooperation.

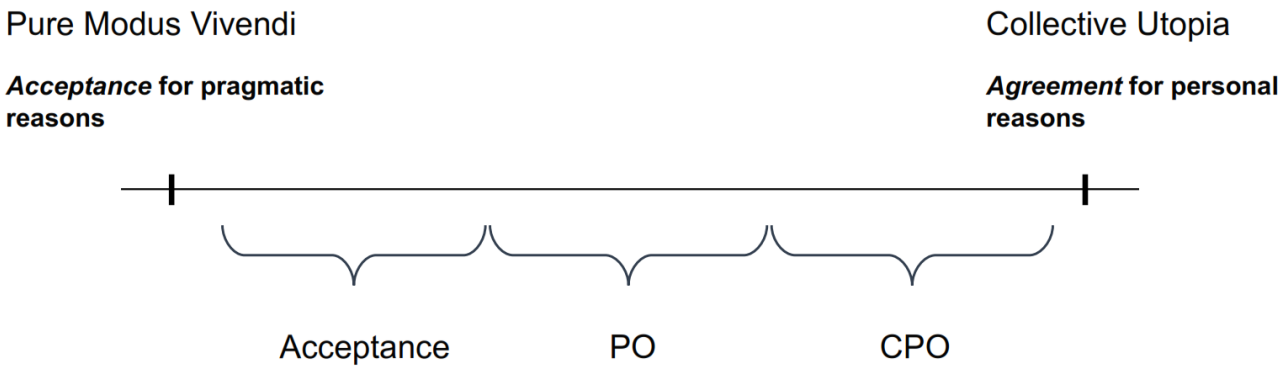
As already pointed out in the previous subsection, the account of justified social order as a compromise with ownership builds on what you might call "the ownership endorsement hypothesis": If reflected upon by some individual, significant coherence between social order and personal reasons may cause that individual to identify with and thus have ownership attachment to said order.

Assuming that this is correct, we can now turn to the elaboration of a systematic account of justified social order as a compromise with ownership. We do this in terms of a hierarchy from the least justifying to the most justifying relation between individuals and their social order. Consider illustration *Figure 4.4*.

Pragmatic Acceptance

On the lower end of the spectrum, but well within the eligible set of justified orders, we find a scenario where individuals favor their social order in virtue of mainly pragmatic reasons. Such individuals strictly prefer the given order to the state of nature and none of its norms are outright unacceptable to them. Nevertheless, the given order does not cohere significantly with

Figure 4.4: Range of Justification III



what such individuals want for themselves personally. Thus, they will probably see their orders as an imposed order, rather than their own order. This is not only an undesirable experience, but also has several functional disadvantages. Most generally put, individuals in this scenario are likely to behave like predicted by classic rational choice theory because they do not have any intrinsic motivation to adhere to the norms they live by.[18] Therefore, they will be likely to disregard the norms when they think they can get away with it.

In organizational science, individuals with such inclinations are sometimes referred to as "good actors"; i.e. individuals who give the impression of being highly cooperative, but usually opt for free riding if it appears to be more beneficial for them in some instance. (Bolino 1999; Griep, Wingate, and Brys 2017: 92) My preferred terminology for individuals with mainly pragmatic attachments to social order is "opportunists". This choice of words is meant to empathize the instability and inefficiency that results from mainly pragmatic attachments to social order. That is, opportunists do not have any motivation to invest effort in upholding social order or to give weight to the spirit of the norms they live by. Further, much like an international cooperation, they are likely to abandon some social order if they discover a more beneficial alternative.

A Compromise With PO

Next in line we have individuals with psychological ownership (PO) for their social order. In such cases, individuals have sufficient personal reasons to identify with the order in question and thus develop ties of psychological ownership. Now, as I have already alluded to above, I cannot offer a general specification of what counts as having "sufficient personal reasons" because this

[18] Assuming that these norms are not already internalized by the individuals in question.

obviously depends on the circumstances and on the personalities of the individuals involved.[19] So for instance, having clear norms itself may be more or less favored by different people. Some people might be always on the lookout for concrete norms to guide their behavior in a given social context, whereas others feel constrained by the presence of norms and rather seek unregulated space to be creative and improvise. The former kind of people will probably develop an ownership relation with norms more easily, whereby this will be more difficult in respect to more free-spirited specimens.

The crucial point on my account is that people can identify with their social order. This brings us significantly closer to an ideal state of justified social order in two respects. Firstly, having ownership of norms through reasoning about norms enables us to live by norms that we can endorse as our own norms according to our own standards. They are of course still unlikely to be our most preferred norms, but this need not be a problem for someone who can nevertheless identify with a compromise as something that is of inherent value to her as the person she is. Secondly, having ownership of norms also means that we have intrinsic motivational resources to follow, enforce and improve these norms. So, norm followers with ownership are likely to be genuine norm followers who behave as if they had internalized the norm. More generally, individuals with ownership are potentially willing to invest extra effort into improving what they perceive to be "my" order. Such behaviors that go beyond what is instrumentally rational are referred to as "citizenship behavior" in organizational science.[20] And, unsurprisingly, psychological ownership is indeed positively correlated to organizational citizenship. (Avey et al. 2009; H. Ozler, Yilmaz, and D. Ozler 2008; O'Driscoll, Pierce, and Coghlan 2006; Van Dyne and Pierce 2004)

Individuals who display such citizenship behavior are sometimes referred to as "good soldiers". (Bolino 1999; Griep, Wingate, and Brys 2017: 92) However, I prefer speaking of *stakeholders*. 'Stakeholders' is less militaristic and emphasizes that individuals with ownership for their social order see it as being of inherent value to themselves. Thus they are motivated to uphold, defend and improve an order they have a stake in. This is of course not to be confused with being a stakeholder in the economic sense, e.g. being a "sharholder" in an enterprise. The stakeholder with ownership attachments to social order has a different kind of stake. It is not an economic investment, but an identification with norms that cohere significantly with what she wants for herself.

[19] Jon Pierce et al. (2003) argue that certain individual differences, such as the strength of innate motives, personality traits and personal values, will influence PO.

[20] "These are behaviors that are consciously engaged in that contribute to or are intended to contribute to an organization's well-being (for example, helping, whistle-blowing, criticizing the status quo, offering suggestions). [...] In other words, they are behaviors that are intended (or are perceived) to benefit others and not the actor per se." (Pierce and Jussila 2011: 104)

In summary, stakeholders, in contrast to opportunists, identify with social order as "my" order. Thus, all stakeholders have a common stake in the norms they live by, such that they will not easily abandon these norms and can be called upon to invest effort into upholding or improving their own order. This disposition to display "citizenship behavior" is of course crucial for social orders – such as democratic orders – that rely on participation and voluntary adherence to norms.

A Compromise With CPO

Third and last, we have individuals with collective psychological ownership (CPO) of their social order. Now, just as with PO, CPO is due to significant coherence of personal reasons and the order in question. What makes CPO unique, however, is that it further requires the individuals involved to share a group identity and this in turn allows them to change their frame of references from the individual self to the collective self:

> "Ownership can be experienced not only on the personal level but also on the collective or group level. Our self-concepts are inextricably linked to the groups to which we belong and vice versa. For SCT [Self-categorization theory], the process of depersonalization implies a redefinition of the self: from thinking in terms of personal identity ("I") to thinking about the self in terms of group identity ("we"). Through depersonalization the group becomes the (temporary) measure of things, and the values and norms that guide our behavior are those of the group with which we (momentarily) identify."
>
> (Verkuyten and Martinovic 2017: 1025)

Jon Pierce et al. (2010: 812-813) suggest that the following three-step process can lead to CPO:

1) The object of PO becomes grounded psychologically; it becomes "mine" for the individual as the individual finds herself present in them, and they become a part of the extended self.

2) The individual recognizes that not only is she psychologically tied to the object, but so are others. Thus, there is a shift in her personal reference from the self to the group and the inclusion of others.

3) Interactive dynamics (i.e., verbal and non-verbal language) create an emergent property that is more than the sum of the individual attributes; whereby agreement among team members emerges, and the construct is transformed from the individual level to the

group level and that collective cognitive/affective state according to which the target of ownership is "ours" emerges.

Getting back to Cillian McBride (2015) and his hypothesis that public deliberation produces ownership, it is obvious to suggest that it is the process of public deliberation that can facilitate steps 1) through 3). Therefore, I believe it is quite plausible that public deliberation can actually produce PO and CPO. Nevertheless, an empirical examination of the relationship is still lacking.[21]

While waiting for the results, let us consider the two advantages of CPO respective to PO, which both build on the fact that it allows individuals to use the group as a frame of reference instead of their individual selves. Firstly, CPO brings individuals closer to being self-legislators. Recall that any individual with ownership enjoys the benefit of feeling as if she is living by her own norms, rather than living according to norms that are dictated to her by somebody else. However, even if someone has individual PO for some order and thus can endorse it as her own order, she still has to live with the fact that she alone cannot change norms and thus is usually not the legislator of the norms she lives by. With the shift from self-reference to group reference this can change because the group as a whole can be the legislator of its norms. So if an individual identifies with the group that, as a whole, understands itself as the legislator, she will be able to understand herself as living by self-legislated norms. This is of course a demanding goal to achieve in social reality. It would require regular citizens to think that the principle of popular sovereignty is indeed well-reflected in their political system and not just a story some elite – i.e. a group they are not a part of and do not identify with – is telling them in order to forge ownership for norms that are really just dictated to them. But besides it being a demanding goal, a sense of self-legislation that is not an outright illusion or manipulation does seem possible in principle via CPO for shared norms. And it sure would be nice to not only live by norms one can endorse as one's own, but also by norms which one perceives as self-selected norms. I believe this is as close as we can get to being a self-legislator in a large and diverse society without being the dictator ourselves.

Secondly, with CPO a group may establish higher forms of cooperation. Recall that in the case of modus vivendi order and social order with individual PO we should expect individuals to focus on their own benefit when evaluating norms. The difference between the two scenarios is merely that in the case of individual PO, people see the norms in question to be of inherent value. Nevertheless, the only relevant agent that evaluates is still the individual self. Now, one can of course simply assume that people are strongly altruistic and consider the costs and

[21] With the established methods of experiments in deliberative mini-publics, which I discuss briefly in *Subsection5.1.2*, and measures of psychological ownership (Van Dyne and Pierce 2004), this research gap could be filled.

benefits of everybody governed by the norms in question, but our framework and our empirical model do not give us any reason to make this assumption. With CPO, however, individuals shift their frame of reference from the individual to the group level.[22] This change has the effect that individuals can be expected to be highly responsive to improvements in overall group benefits while – and this is crucial – this does not have to translate into individual benefit.

The most obvious advantage of a state where people replace group accounting in favor of individual accounting is that this should make it fairly easy to coordinate on the provision of public goods according to a solidarity principle. An example here would be universal health care that is provided in a way that everybody is required to contribute to the costs according to her capabilities, whereby everybody benefits from it according to her needs. In such a scenario, individuals with high capacities for contribution but few needs are likely to pay more into the system than they get out of it. So if this is not the kind of person who values solidarity as an end in itself, she will be unlikely to contribute voluntarily. This might change, however, if her frame of reference is the group at large, given that public health care is overall beneficial from the group perspective. Generally speaking, individuals with CPO are accessible to public good consideration. This is of course not to say that CPO is the only way to get to the provision of solidary public goods. Internalized norms or political institutions may also work. However, sticking to our example, solidarity may not be a shared norm in a diverse society. Consequently, implementing institutions such as a state-backed health care in a solidary and universal way may be costly and politically unstable. Thus, CPO offers an attractive alternative for establishing things that are overall beneficial from the group perspective, but not necessarily beneficial to every individual.

Within organizational science there is, to my knowledge, no metaphorical category, such as "good solider" or "good actor", specifically for individuals with CPO attachments. My suggestion here is to speak of individuals with CPO as "citoyens". The citoyen is characterized by identifying herself as a part of a community of citizens and thus she has a unique motive for being concerned with the general interest of her community.[23] In a way, citoyens are like evolved stakeholders. They are like stakeholders in that they are motivated to take responsibility for *their* social order. Further, and in contrast to stakeholders, citoyens identify as members of a community of collective owners. Thus, whereas the stakeholders might pursue a reform that first and foremost benefits themselves, the citoyens pursue reforms that are in the group interest – in "our" interest.

[22] "Depersonalization redefines self-related terms: It is about collective self-esteem, collective self-efficacy, and collective self-interests rather than personal self-esteem, personal self-efficacy, and personal self-interests." (Verkuyten and Martinovic 2017: 1025)

[23] Speaking of "citoyens" is of course also a nod to how this concept and the idea of a general will is presented by Jean-Jacques Rousseau (1762: Book I, Chap. VI).

Overall, citoyens, identify with the norms they live in terms of a collective of owners – a "we". With this social dimension added, we might say that a community of citoyens, who, for the right reasons, share a group identity and a deep concern for "our" norms, have a social contract.

This concludes the normative model. In *Subsection 3.2.3* I introduced the ideal of Political Equilibrium in order to demarcate a certain perspective on and potential theory of justified social order. This theory seeks an open-ended, procedural ideal of how a group of individuals could be thought of as being governed by a set of stable social norms for the right reasons. At this point, we have finally arrived at a specification of such an ideal: Ideally, social order can be a compromise with ownership, i.e. a compromise that is endorsed by a group of individuals as their own order in light of the pragmatic and personal reasons they have.
Most ideally and most optimistically, social order is a compromise with collective ownership. That is, individuals have their own personal reasons for endorsing social order *and* they share a group identity, allowing them to see it as "our" order. Such a community of citoyens can perceive of themselves as self-legislators of their social order and coordinate on advanced forms of cooperation.

4.3 Discussion of the Ideal

We now move on to a discuss of some critical aspects and particularities of the normative model. More specifically we will attend to the following three questions: How do we get from reasoning to ownership? How do we avoid the dangers associated with ownership and group identity? How do we get from the model view to a more realistic picture?

4.3.1 Reasoning and the Routes to Ownership

We begin our discussion with a reflection of the core empirical claim evolved in the introduction of psychological ownership to our normative model, namely that reasoning about norms can lead to psychological ownership of these norms. This point is crucial because it is a central node of the normative model, holding together its descriptive, functional and normative components. If this claim turned out to be false, we would either have to replace psychological ownership with another construct of psychological attachment that can serve the same function, or abandon the strategy of a psychological specification of different states of justification altogether. As we have seen above for Cillian McBride (2015) and Cristina Lafont (2020) it seems obvious that public deliberation produces ownership. But as it stands, this hypothesis lacks systematic empirical inquiry. However, for the following two reasons I think we can be at least optimistic that the relationship between reasoning and psychological ownership holds.

The Routes to Ownership

The first reason is that reasoning is a plausible cause of ownership attachments. To see this, consider what Jon Pierce and Liro Jussila (2011) identify as the three "routes" to PO. The first route is having control over the object of ownership. Essentially, "the more we believe that we control and influence an object, the more we possess the object, and the more it becomes a part of the self." (Pierce and Jussila 2011: 78) Having control over something attends the basic psychological need to feel effective or powerful in one's environment. The second route to ownership is coming to know the object of ownership intimately. This is to be thought of as a continuous interaction (investigation, communication etc.) with the object of ownership. The third route is investment of the self into the target of ownership. This typically happens through working with or on the object and thereby also manipulating it according to one's desires. Effectively, an activity such as working on some object might send one down all three routes to ownership simultaneously.

Out of these three routes, intimate knowledge seems to be the most obvious candidate for understanding how reasoning about a norm could lead to ownership ties with that norm. That is, reasoning about a norm seems to be an important way to gain intimate knowledge about it, because as we reflect upon the purpose of a norm, possible alternatives and the respective reasons for and against it, we can gain a comprehensive understanding of what it is all about. The other two routes – control and self-investment – seem to be less obvious candidates for relating reasoning and psychological ownership because, more often than not, the norms we live by are beyond our control and created by others. But perhaps reasoning about norms is actually a mental way around the divide between us and the norms produced by others. To see this, consider that a norm is mostly a mental entity, like an idea with normative content, in contrast to a non-mental entity such as a table. Now, in the case of a table, it is probable that the person who has built the table and keeps it in her workshop, i.e. the person who has effective control over and invested effort into making the table, considers it "her" table. Another person, who merely sees the table through the window of the workshop, does not. Assuming that the latter person has no past experience with the table or its maker, it would require a lot of imagination, and effectively self-delusion, to make herself believe that this table was somehow "hers". But with mental entities that are like ideas, things are different because they are built and manipulated in our mental workshops – our minds – and different people can construct and manipulate some idea, irrespective of others coming up with the same idea. So in the case of reasoning and norms, it is plausible that even reasoning oneself to already existing norms is more like creating them for oneself, much like creating an idea, even though someone else already had the idea and has written a whole book about it. Admittedly, only the content of a norm can be like an idea. The norm itself, if it is an active social norm, is

part of a shared social reality that is usually somewhat independent of the individual mind. At the same time, and in contrast to the table, the existence of the social norm depends in its existence on each individual mind. Therefore, I do think it is plausible that, by reasonably recreating a norm X in our own mind and making practical decisions to follow or not to follow the norm based on our respective reasoning, we can have a similar sense of intimate knowledge, control and self-investment as the carpenter making the table. And as we observe X being followed by others in our community, we observe our own reasoning reflected in the social world surrounding us and it may very well be at this point that we identify with that social norm.

Reason and Identification

A second reason in support of the hypothesis that reasoning leads to ownership is the nature of reasons and what they do for us in practical deliberation. What reasons are is a heavily contested matter in metaethics. Nevertheless, we should be able to agree that reasons in practical deliberation show certain actions or states of the world to be favored (or disfavored) relative to some given agent. The last part is crucial: reasons favor something relative to some agent experiencing the favoring, which explains why this agent feels a normative force to do what is favored or avoid what is disfavored. The normative force of favoring, experienced by some individual, is what you might call the "for-me" character of the favoring done by reasons. (Fischer 2018: 22; 60-61) It is the experience that occurs to you when you walk past a bakery and remember that you do not have anything to eat at home and feel an urge to go inside to buy a loaf of bread. My suggestion in respect to reasoning and norms is that reasoning about norms and discovering personal reasons favoring these norms could, especially due to the for-me character of reasoning, be the basis for the kind of identification with a norm that can lead to psychological ownership. More simply put, the idea is that realizing that I have personal reasons in favor of some norm allows me to see a part of myself within that norm. This in turn can lead to self-identification with the norm in question and, consequently, to a sensation of psychological ownership.

Now, irrespective of how powerful you believe these points to be, there remains a speculative gap. That is, we do not know for sure whether psychological ownership, internalization into the norm system, or something else explains how a group of individuals can get from reasoning to effective norm following. Which explanation is the correct one remains to be established by means of appropriate experiments. Perhaps they are both true.

Be that as it may, the distinctive feature of psychological ownership is that it can explain the pro-active "citizenship" behavior of people I have characterized as stakeholders and citoyens. More specifically, ownership attachments can explain why certain individuals may be motivated

to display individual or collective efforts to improve *their* social order. This pro-active stance is crucial for the practical plausibility of Political Equilibrium, especially in respect to the idea of actual public deliberation about norms, which we will discuss in detail in *Chapter 5.* Also, democracy as we know it would be hard to imagine without pro-active citizens participating in the advancement of the system.
In spite of all that we have discussed so far, there remains much to be clarified. In particular, the relationship between psychological ownership and reasoning as well as between the different psychological constructs of PO, CPO and citizenship behavior need to be investigated further – not only within organizational science, but in social psychology and social science more generally.

4.3.2 The Dark Side of Ownership

It is a peculiar feature of our normative model that the ideal state it specifies has two core components – psychological ownership and group identity – that are independent of its core normative qualification – justification as coherence between well-considered individual reasons and norms. Consequently, psychological ownership and group identity can exist without being normatively qualified in any way.[24] This is important to consider because both states are not only associated with effects facilitating, but also hindering cooperation.

The Dangers of PO

Firstly, consider individual PO. It has been argued that PO can be accompanied with a reluctance to share. (G. Brown and Robinson 2007) For instance, it might be the case that some individual is not sharing knowledge that would be of value to others or even to the whole group. (Peng 2013) The general problem here seems to be that individuals consider something to be *exclusively* theirs. This kind of exclusive attitude, however, is not a realistic scenario for ownership of *social* norms. Social norms, per definition, always relate to a group of individuals and it would be odd if one of them considered them to be exclusively her norms. If this were to happen, such a person would miss something very fundamental about the purpose and workings of social norms and she would probably suffer some social-psychological dysfunction. Social norms always apply to some group of individuals and while they need not have any group identity (think of people on a public bus), they should be aware of the fact that the norms are not in place exclusively for them as individuals. In short, individual PO for norms can only be non-exclusive ownership. In the case of CPO, however, things are different. We

[24] "Citizens may, however, regard themselves as standing in a relationship of ownership to their political institutions without good reason. This may be adequate for producing a measure of political stability, but these cases should be just as worrying to the democrat as cases of political alienation." (McBride 2015: 110)

will get to that shortly.

Another problem with PO raised in literature is that PO alone does not motivate engagement in group efforts, whereas CPO does. (Martinaityte, Unsworth, and Sacramento 2020) This is coherent with my above claim that below the level of CPO, individuals remain maximizers of personal benefit and thus are vulnerable to collective action problems (e.g. lack of trust). In absence of CPO, these problems will need to be addressed by an additional formal or informal enforcement regime. This is one important reason why I consider CPO to be more desirable. Lastly, in respect to PO there is a tendency to be blind to critical knowledge and change in respect to what is perceived to be "mine". (Baer and G. Brown 2012) This is an important point to keep in mind for the design of the ideal social procedure in *Section 5.1.* Generally, I assume that our insistence on an open-ended and procedural ideal and its proper practical realization will address this issue to a large extent. That is, an actual deliberative, political process that reflects this ideal should go a long way toward preventing excessive conservative biases because such a process would be characterized by continuous critical examination, discussion, revision and adaptation.

The Dangers of CPO

Secondly, let us critically reflect on CPO. I have said in respect to PO for norms that exclusive ownership is not likely to occur. This is of course not the case with CPO. CPO for a group's norms may very well be exclusive. Thus CPO may lead to an exclusionary stance toward non-members of the group and inter group tensions between different groups. (Nijs et al. 2020; Verkuyten and Martinovic 2017)

Although exclusionary ownership is indeed a possible side effect of CPO, it does not result from an individual's well-reasoned identification with a set of norms, but rather from the nature of the group identity and the ownership relation. To clarify these two points, let us begin by considering the extreme but illustrative case of a society of racist nationalists. If these individuals would cohere significantly in their nationalistic views and their racism, they could probably achieve CPO in respect to a social order that is highly exclusive, especially toward outsiders of a different ethnicity. Now, from the standpoint of justification, we have to acknowledge that this kind of order is probably highly justified to those kind of individuals. We may of course intervene argumentatively by pointing to empirically false claims that many racist stances are based on or point to the fact that communities that are more open and cooperative in respect to outsiders and other communities also do better for themselves. But these arguments might not succeed because racism is usually, but not necessarily, based on false assumptions. And because for our group of racist nationalists, the "purety" and autonomy of

their community might trump all other considerations. So a highly exclusive, racist social order can be highly justified relative to a homogeneous group of racist nationalists. Fortunately, large and diverse societies are not homogeneous in this way.

Now, also consider that things could be quite the opposite. Here you might think of a society made up of cosmopolitan democrats. Given that these individuals also cohere significantly in their understanding of cosmopolitanism and democracy, they may achieve CPO in respect to a highly inclusive social order. That is, they could identify as a group of people who constitute a community of shared inclusive values (e.g. universal human rights, including democratic rule) and a social order reflecting these values. This order would probably explicitly allow for and manage the inclusion of previous non-members and generally members will look at their order as something that should be expanded to include (or "benefit") other people. In that sense, our cosmopolitan democrats will be *missionaries* of their social order, while the racist nationalists will rather be *gate keepers*.

The point I am trying to make with these examples is that CPO for rules may be inclusive, just as it may be exclusive. It all depends on the (cultural) programming of the ownership relation and group identity.

Another challenge for the inclusive and diverse society in respect to group identity is this: How is it possible that we identify with others who we do not know and who are probably very different from us in many respects? The answer seems to be that this is indeed possible if we construct group identity in the right way. To see what I mean, first note that there is no contradiction between group identity and individualism. Rather, there are many ways of expressing individualism and fulfilling one's need for being different from the group in coherence with group identity.[25] Secondly, note that the creation of group identity can be a dynamic "deductive" *and* "inductive" process. That is to say that on the one hand, individuals may deduce group properties, i.e. stereotypes and group norms, top-down, from the group to the self. On the other hand, group identity may be shaped inductively, bottom-up, by the interaction and communication of individuals. Essentially, in order to accommodate diversity and change, there should be a continuous, circular dynamic between individuals and their group identity. (Postmes and Spears 2005; Postmes, Baray, et al. 2006)

Combined, these two points clarify that group identity is not necessarily an inflexible doctrine fixated on characteristics all group members have to share. Quite the contrary, group identity

[25] "[D]istinctiveness needs can be met through (rather than despite of) group identification. By (a) identifying with a numerically distinct group, (b) identifying with a subgroup, (c) identifying with a group that defines itself against the mainstream, or (d) perceptually emphasizing the distinctiveness of one's group, participants can feel the comfort of belonging and inclusiveness without sacrificing their need for distinctiveness." (Hornsey and Jetten 2004: 254)

can explicitly embrace diversity. (Cunningham 2005) If this is the case, the group may even "interpret a display of distinctiveness as a sign of trust in the collective on the part of the deviant." (Postmes and Spears 2005: 749) Unfortunately, there are many examples in human history for group identities that have mainly been handed down from the group to the individual by means of seductive ideologies. These ideologies have told people that they belong to a unique group (the "Aryans", the "proletarians", "The chosen ones" and so forth) and will be well off, secure and proud if they are only loyal to the political regime claiming to represent this group. These ideologies are "seductive" because they offer easy ways for group identification by means of exclusive in-groups, pointers to out-group enemies and rituals (e.g. parades, and symbols of party and subgroup memberships) which devise a sense of participation and belonging without offering any real political participation.

It seems likely that this kind of top-down group identification is mainly responsible for dangerous orders with CPO that are highly exclusive and do not place any value in individual, but only in group ends. As the cited research shows, however, that diversity and bottom-up construction can be a core feature of group identity as well. Thus, a diverse group, where individuals critically reflect upon the nature of the group's identity and its norms, is possible. What is required, however, is that the (diverse) individuals are able to achieve common cognition and group identification during their discussion in order to successfully negotiate cooperative outcomes. (Swaab et al. 2007) One practical suggestion in this context is that diversity should be built into our very conceptions of citizenship and this should be reflected in citizenship education. (Banks 2008)

The concept of democratic citizenship is indeed interesting here because it offers an elegant way of aligning ownership for norms with a a group identity that is compatible with diversity and bottom-up alteration. In detail, the group identity of democrats has itself a certain social order as its content. Thus we would expect members of this group to converge on personal reasons in favor of a shared democratic order and a shared identity as proponents of this order. This would ensure that individual reasons for valuing group membership and for valuing group norms are unlikely to come apart.

Further, due to the nature of the typical kind of procedures and rights that come with democratic social order, a group identity reflecting this kind of order is very unlikely to be exclusive or only a top-down construction. However, no matter how elegant this scenario might appear, it does not need to be the only way for a diverse community to achieve justified social order with CPO. One reason for remaining agnostic here is that the relationship between group identity and ownership of group norms is difficult to clarify. And I am not aware of a systematic, empirical inquiry into the matter. Therefore, at this point we are well advised to leave some room for further investigation.

Another problem with CPO for norms is that group identity may lead to a horrific scenario of a dominant group identity where individuals may be treated as means for group ends in an unrestricted or marginally restricted way. As a consequence, individual lives and concerns are of insignificant value and may easily be sacrificed for whatever is presented as the greater good of the group in question.[26]

I believe the danger of a dominant group identity is also largely addressed by the requirement of a critical, individual evaluation in public deliberation. Recall that, according to our normative model, the way for each individual toward ownership for norms is her identification with these norms based on the coherence of her well-considered reasons and the content of those norms. And, although I have bracketed the issue in the normative model, recall that I have also said that well-considered reason and preferences are practically only conceivable as the outcome of public deliberation. This possibility for critical evaluation in public deliberation should set a limit on the degree of group dominance, defined by the very set of personal reasons that allowed for the ownership relation to develop in the first place. Simply put, the idea is that if you learn that the group action violates the values or ends that directly motivate your identification with the group, or other important personal values or ends which had not been effected yet, you might lose your ownership attachment. Here you could think of your government declaring a war to fight some injustice, but you have always been deeply convinced that wars cause the greatest injustices.

However, there is a practical problem here. Namely that the things we value, the goals we have and, all in all, the kind of persons we are, is largely a function of the particular culture and social order we have been born into. This is a classic point communitarians raise against liberal views of autonomy and a self-chosen life. (Bell 2020) To clarify, the problem is not that we are socialized into a give social order. It is normal and perhaps necessary that the justification of some social order is the product of some cultural-historic development that also created the order itself, and through it, keeps reproducing a kind of circular justification. Simply put, democracy is well-justified to and works through a democratic citizenry, whereby democracy itself reproduces this kind of citizenry. This circularity is difficult to escape for societies and socialized individuals. But we do not need to escape our history or our socialization. The crucial point is that *partly* self-regulating individuals can and do gain a critical perspective on the social orders they are born into and test whether the justifications passed down through history or their peers still holds for them. Establishing such a critical perspective should be a core task of the social mechanism producing PE. We will return to this point in *Chapter 5*.

The real danger here consists in the possibility that a group of individuals achieves something

[26] Essentially these are the standard problems of group utilitarianism that have motivated John Rawls' and Robert Nozick's concern for the separateness of persons. (Rawls 1971: 27; Nozick 1974: 32-33)

that very much looks like justified social order, while it is really not justified at all. This can happen when the critical aspect of public reflection does not really materialize. A rather blunt version of this problem would be a political regime that intentionally tries to prevent critical thinking in order to preserve their own power. This however is often fairly obvious and we know the kind of norms needed to prevent it: free speech, freedom of the press, absence of censorship and so on. A more subtle version of the problem occurs when a group is locked into a kind of "groupthink" that ignores critical (minority) views, new information and alternative ways of organizing social life. For instance, and of particular relevance to our normative model, Robert Baron (2005) argues that group identification, producing or revealing group norms and low self-efficacy in light of a complex problem are jointly sufficient for producing groupthink phenomena. These are the kind of problems we need to be aware of when designing a mechanism of collective reasoning.

So far I have presented collective reasoning mainly as a remedy for the faults of individual reasoning. But of course, also the former comes with its own kinds of problems. Thus any attempt of specifying the ideal social mechanism of PE will have to pay close attention to the challenges involved in collective reasoning.[27]

4.3.3 From the Model to Reality

As we exit the normative model and begin to grapple with more practical questions, many things will get more complex. One example that we have just considered is the right calibration of public deliberation so that it actually provides us with a critical evaluation of ourselves and our social order. I have mainly bracketed this issue in the normative model, but it will take center stage in *Chapter 5*. However, there are more complexities to consider.

Relational Chaos

Another aspect in terms of getting a more realistic picture are the different reasons that individuals will have in favor and against some existing social order. Our discussion of empirical and normative *models* in this chapter has led us to think about the social world in terms fairly simplistic and clearly arranged ideas and conceptions. But of course, any real world scenario will be characterized by a chaos of different relations between individuals and their social order. To see this, consider on the one hand that, realistically, citizens are neither all opportunists, stakeholders or citoyens. They are just theoretical archetypes. Real citizens may embody all three of them to different extents and depending on circumstances, triggers and framings.

[27] Julian Müller (2019) provides a helpful overview of the different kinds of defects and biases that can distort group discussions. We will briefly discuss his analysis a the end of *Subsection 5.1.4*.

Further, individuals may be guided by things, such as internalized behaviors, that are not captured by the normative model.

On the other hand, relations between norms and individuals may vary significantly. That is, there is probably a colorful mixture of norms that are pragmatically accepted by some, but personally endorsed by others. Within the latter group, there may be individuals who only have individual ownership and those that see themselves as part of a community of shared values and corresponding norms. There may even still be those who consider some norms to be outright unacceptable. All of these relations may exist simultaneously between different individuals in respect to some norm and also between different norms and the same individuals. In spite of all this chaos in the real world, our modeled ideal might provide us with some guidance here, by suggesting how to react to the different relations. The most obvious suggestion is to try to eliminate norms that are outright unacceptable. This is of course problematic if this elimination is unacceptable for another group of people. Conflicts on legal abortion and gay marriage can exemplify this problem insofar as they are conflicts of conflicting personal values. I do not have any solution to offer for cases of value-induced division. However there are some hopeful examples that such division may be overcome through public deliberation.[28]

Besides the concern for the focus on norms that are outright unacceptable or produce conflicts of personal values, basically what the normative model suggests is that we focus on reforms that allow people to move up into higher stages of justification: That is, turning opportunists into stakeholders and stakeholders into citoyens. The reason being that this will allow people to feel more autonomous and be more motivated to contribute to the implementation and improvement of their common order. Overall, more people coming closer to the level of CPO will produce a rather cohesive and participatory, instead of an alienated and apathetic citizenry. To clarify, the suggestion here is not to maximize the satisfaction of personal reasons across the board, but to seek reforms with the potential to include more individuals in higher stages of justification without dropping others. This could be done by adding to the values reflected in a given social order, as long as this does not produce new defeater norms, or by extending the range of existing rights and freedoms to previously excluded parties. If inclusion does not help, polycentric order (i.e. allowing for diverse, partial sub-order) may be a way forward. We will briefly discuss this option in *Subsection 5.1.4*. If this also does not work, separation may be the only option that is left. As I have said before, justified social order is not guaranteed for

[28] See for instance the anthology by Juan Ugarriza and Didier Caluwaerts (2014): *Deliberation in Deeply Divided Societies: From Conflict to Common Ground*. London. A particularly interesting contribution in this volume is Robert Luskin et al. (2014), who have conducted a deliberative poll in Northern Ireland between Catholics and Protestants about the future of local schools. Not only did this fairly short intervention have a measurable effect, it also uncovered that ordinary citizens are often less divided than grand narratives of a divided society might have us believe, and having people discussing practical matters is an effective way of dismantling this destructive illusion.

every combination of individuals, even more so if we are considering its most ideal incarnation of justified social order as a compromise with CPO.

In this section we have discussed several critical points regarding the arguments and outcomes of the normative model. All of this was meant to address some of the most imminent worries in relation to PE as specified in the normative model (I summarize the main points in items 9-11 of the concluding summary right below). I hope that the discussion at least establishes that the specified ideal is worth pursuing in more practical terms, which is what we will turn to in *Chapter 5*.

4.4 Concluding Remarks *Chapter 4*

The goal of this chapter was to work out Political Equilibrium as a theory of justified social order in more detail and thereby to showcase the methodological approach I call Embedded Constructivism. To this end we have, first, specified the underlying empirical model and, second, explicated a suitable normative model. Thus we have completed the first two steps of Embedded Constructivism. Third, we have taken the time to discuss some implications and peculiarities of the normative model. Here is a summary of what has been argued:

1) The emerging theory of social order is fairly clear about what facilitates stable cooperative order, understood as a set of social norms: A community made up of a majority of rule followers, possibly some defectors and some rule-following punishers who punish the defectors.

2) The emerging theory of social order is less clear about coordinating on a particular cooperative order. Particularly the psychological relation between internalized norms and reasoning about norms remains uncertain. What we do know is that norms can be changed by changing normative and empirical expectations, i.e. by reasoning for *and* acting upon new norms.

3) The core image of the empirical model is that of partly self-regulating individuals who seek to coordinate on cooperative norms while having lost their biological standard of mutual benefit in reproductive fitness for solving their coordination problem. Hence they are left with the problem of choosing from a set of eligible, cooperative norms.

4) Although being almost implied by the fact of social order and diversity, the first leap from the empirical to the normative model is the assumption that individuals are pragmatic enough to accept social order as a mutually beneficial compromise: In face of persisting

disagreement on the best norms, they settle for a compromise, which is mutually beneficial in the eyes of each individual.

5) The normative model starts out with the insight that eligible and justifiable compromises are likely to be located well between the minimum point of being preferred to the state of nature and the maximum point of a comprehensive agreement on the best norms. Further, the compromise should not include outright unacceptable norms, which defeat pragmatic, overall acceptance.

6) The second main step in the normative model is the introduction of three qualitatively different relations between individuals and their social order: One, acceptance of a social order for primarily pragmatic reasons. Two, endorsement of a social order for primarily personal reasons, facilitating psychological ownership for that order. Three, endorsement of a social order for primarily personal reasons, combined with group identity, facilitating collective psychological ownership for that order.

7) Particularly, justified social order as a compromise with collective psychological ownership has emerged as the most desirable, and still realistic, ideal. It is realistic because we know that groups of actual individuals can achieve this state. It is desirable because with collective ownership individuals are likely to actively participate, reach advanced forms of cooperation and experience their norms as self-legislated.

8) In discussing the outcomes and implications of the normative model, I have explicitly defended the crucial claim that we can reason ourselves to having ownership of norms. The defense rests on the points that, one, reasoning about norms is likely to send individuals down the "routes" to psychological ownership. Two, reasons as such are well-suited means for identifying with something.

9) I further discussed the danger of ending up with group identities and ownership relations that contradict having a diverse society with a justified social order. Here I argued that the normative background condition of well-reasoned endorsement as well as the right programming of ownership relations and group identity (e.g. making diversity part of the group identity) can in principle address most problems.

10) Lastly, I discussed the differences in the relations between individuals and their social orders. Here I pointed to the potential of the normative model in offering general guidance on reforms: Avoiding defeater norms and choosing inclusive reforms that realize potentials for more ownership.

Before we move onto the last chapter, I would like to add a humble note of caution. In order to get to a more substantial account I have introduced a range of psychological constructs. Most importantly, I have introduced the construct of psychological ownership. The basic motivation for doing this is to have an account that can explain why justification and reasoning about norms matters in actual social life. In *The Order of Public Reason*, Gerald Gaus (2011) introduced the idea that people can internalize a norm that is well-justified to them, much to the same end. Perhaps he is right. But as we have seen in the empirical model, the psychological mechanism that allows partly self-regulating beings to deliberately choose their norms is unclear. Perhaps reasoning about norms does offer a psychological route to internalization. Perhaps it does not.

Reasoning ourselves toward ownership for norms is meant as an alternative account of how we get from reasoning about norms to being intrinsically motivated norm followers of self-chosen norms. The advantages of this alternative explored above are, one, that reasoning appears to be a plausible route to ownership attachments. Two, that having ownership offers a broader motivational basis that can explain why people are not only rule followers, but active citizens. In spite of these advantages, our normative model makes a range of assumptions which are still in need of examination. This includes the thesis that reasoning about norms can indeed lead to identification with a norm if the norm coheres significantly with the personal reasons on has. It also includes the assumption that the different constructs of psychological ownership and organizational citizenship are replicable outside of organizational science. Further, even within organizational science, authors such as Yannick Griep et al. (2017) have only begun to investigate whether these constructs integrate well, as I have suggested in the normative model.

In conclusion, there are several crucial aspects of the normative model that require empirical scrutiny. As we gain more knowledge about these things, the model might require significant refinement. This, however, is the normal working mode of Embedded Constructivism where normative theorizing builds on an empirical basis: As our understanding of this basis advances, so do our respective concepts and normative models.

Now we move on to three important more practical issues the normative model has not addressed: Firstly, the question as to how model citizens can be thought of as actually having well-considered preferences on social order (Condition 1 of the JPN). In the normative model, we have simply assumed that model individuals have well-considered preferences, but of course, this assumption is questionable with respect to actual citizens. Secondly, we also need to explain how citizens themselves can critically assess the theory of Political Equilibrium (Condition 3 of the JPN). Thirdly, the theory of Political Equilibrium presented so far is still lacking a

crucial aspect of Embedded Constructivism: relating the open-ended, procedural ideal explicated in the normative model back to social reality. These are the three tasks of the following chapter.

Chapter 5

Of Mechanisms and Tests

The core result of the normative model constructed in the previous chapter is a fairly specific understanding of Political Equilibrium (PE): justified social order as a compromise with ownership. The challenge we are now facing consists in explaining the practical relevance and meaning of this ideal. As already worked out in *Chapter 3*, I have two complementary strategies in mind for approaching this crucial last step. The first one, pursued in *Section 1*, consists in thinking about how PE could be translated into a real-world social mechanism for norm selection. The second strategy, pursued in *Section 2*, consists in thinking about a test for establishing to what extent some given society has already realized PE in respect to its order. Crucially, both strategies can be pursued simultaneously and will complement each other in practice. That is, a given society will benefit from an independent test for whether the mechanism is working. Conversely, just having a test would be unsatisfying without a social mechanism that allows for the systematic improvement of one's score.

Overall, the core idea that explains how citizens can reason themselves toward justified social order with ownership is, as Cillian McBride (2015) and Cristina Lafont (2020) have already suggested, public deliberation. Thinking about a society where citizens openly deliberate about what norms they should live by also allows us to explain how Condition 1 and 3 of the JPN can be satisfied.[1] That is, participation in the forum of public reason can explain how actual citizens may come to have well-considered preferences on social order. At the same time, it puts individuals in a suitable position to critically evaluate the plausibility of my theory of justified social order themselves.

[1] Recall the justification principle for social norms (JPN): A social norm N is justified to an individual i in society S governed by that norm to the extent that N being a positive norm in S is coherent with i's preferences, given that (1) i has formed well-considered preferences on social order, (2) N being a positive norm in S is strictly preferred by i to having no social norm governing the domain of N in S, (3) i is at liberty to openly reject the JPN in S.

5.1 Mechanism Design

We begin by reflecting upon the notion of a social mechanism, aiming at the realization of PE. Let us call this mechanism "Political Equilibrium Mechanism" or simply "PEM". Effectively, the PEM should ensure that a society's selection of social norms gravitates toward selecting justified norms with ownership, just as a market mechanism is meant to ensure that selected distributions gravitate toward the state of market equilibrium.
However, as already pointed to in *Subsection 3.2.3*, putting together a comprehensive outline of the PEM mechanism is an enormous task. Therefore, in this subsection we will only be able to take some first steps until coming up against the boundaries of what can be achieved within this inquiry and in normative theorizing more generally. My motive for nevertheless getting into the topic of mechanism design is to show that the Theory of Political Equilibirum (TPE) does connect to more practical matters and theories of norm selection. Overall, I mean to show that there is potential for a symbiotic relationship between TPE and more practical discussions in political theory and science. The symbiosis consists, on the other hand, in that a normative background theory such as TPE can provide guidance to more practical accounts and discussions. On the other hand, a theory such as TPE gains in substance, precision and meaning, as it is being extended to more practical discussions.
We proceed by firstly looking back at all the tasks we have associated with the PEM in the preceding sections. Secondly, we will relate TPE to the literature on deliberative democracy. Thirdly, we will relate TPE to a wider proceduralism that includes considerations of justified social order in a diverse society beyond deliberative democracy.

5.1.1 Tasks at Hand

Throughout this inquiry, we have come up with several tasks to be fulfilled by the PEM. It is time to take stock of these different tasks for they will serve us as a general guide throughout this section.

Well-Considered Preferences: Recall that individuals are unlikely to achieve well-considered preferences on social norms on their own. This is because, as has been pointed out in the empirical model[2], human beings often fail at reasoning well on their own. Also, as elaborated in *Subsection 4.2.1*, an isolated individual is unlikely to have the necessary knowledge and (political) experience to produce well-reflected preferences on norms. Hence, the PEM has to ensure that individuals can engage in a practice of collective reasoning about their norms, which tends to produce well-reflected preferences on these norms.

[2]In particular by Hugo Mercier and Dan Sperber (2017).

Critical Reflection: A crucial aspect of achieving well-considered preferences that needs emphasizing is a critical reflection of the status quo and potential reforms. To this end, preference formation has to be exposed to critical thinking and critical voices. As our discussion in *Subsection 4.3.2* has shown, achieving a critical perspective on the status quo and potential alternatives is what distinguishes justified social order from merely reproducing social order and what protects collective reasoning from biases, groupthink phenomena and manipulation.

Balancing Group Identity and Individuality: PE as specified in our normative model prescribes a social state in which individuals share a group identity while retaining their own (critical) perspectives. Striking such a balance is a demanding ideal. It requires, on the one hand, common goals, values, experiences, narratives and symbols to be actively sought and publicly reproduced. This active facilitation of group identification can be a powerful enabler for collective action and collective goods. On the other hand, it requires individuals to have a protected space for critically evaluating group identities and group decisions as well as for voicing disagreement, criticism or outright rejection.

Selection of Justified Norms: Any actual mechanism will have to include political decision-making in order to select one norm over another, even if the matter remains controversial. The core challenge in doing so, according to TPE, is selecting norms from within the eligible set of justified social order while maximizing ownership. In more detail, norm selection has to cater to certain general constraints and a goal. The constraints consist in avoiding outright unacceptable norms and norms that are not strictly preferred to having no norms in place. The goal consist in identifying reforms that allow more citizens to achieve PO, or even CPO, in respect to their social order.

Self-Justification: According to the JPN any account of justified social order must provide for self-testing.[3] This accommodates the fact that no account of justified social order, including TPE, is beyond reasonable criticism and rejection. More specifically, the construction of the normative model is based on the assumptions that actual citizens value publicly justified social order and having ownership for their order. These assumptions may turn out to be false for particular individuals or groups. Further, TPE does not include trade-off rates for PE versus other things citizens might value. Last but not least, the specifics of a procedure such as PEM do not follow deductively from the ideal specified in the normative model. Eventually, these things come down to political decision-making by actual citizens. Therefore, citizens should not be hindered but rather enabled to critically reflect on the assumptions, conclusions and practical implications of TPE.

[3]See *Condition 3* of the JPN and its discussion in *Subsection 1.3.2*.

5.1.2 Public Deliberation and its Institutionalization

How can the listed tasks be accomplished in a diverse society? Since our ideal is about looking for social order that is stable for the right reasons, the most obvious suggestion for a society of model individuals is that they should try to reason themselves to this state. So, whatever the more specific tools of the PEM may turn out to be, they will have to include a forum of public deliberation.
'Public deliberation' denotes an openly accessible discussion about a topic of common concern, in our case norms or sets of norms, understood as an actual social practice approaching certain ideal conditions. Conceptually this means that the idea of public justification explicated in *Subsection 1.1.2* – a justification given "to all" – in practice translates into justification given "in public". Practically speaking, 'public justification' denotes reasoning that is addressed to all *and* taking place in public – i.e. in the forum of public deliberation. This forum is meant to be perfectly suited to the power of free-floating reasons - an "ideal speech situation". Jürgen Habermas is well known for stating and restating this idea over the last six decades. In a fairly recent contribution he defines the most important conditions for ideal public deliberation as follows:

> "(a) publicity and inclusiveness: no one who could make a relevant contribution concerning a controversial validity claim must be excluded; (b) equal rights to engage in communication: everyone must have the same opportunity to speak to the matter at hand; (c) exclusion of deception and illusion: participants must mean what they say; and (d) absence of coercion: communication must be free from restrictions that prevent the better argument from being raised and determining the outcome of the discussion."
>
> (Habermas 2008: 50)

In Habermas' constructivism, there is a strong empirical claim associated with these conditions. Namely that they are – "de facto" – presuppositions that people have and are guided by when engaging in an argument. This is indeed a strong claim. My conjecture is that public deliberation is a mode of communication with a certain cultural-historical background.[4] Consequently, just like any other cultural practice, public deliberation has to be taught, trained or somehow institutionalized, otherwise it is unlikely to exist in any predispositions of actual reasoners.

[4]Therefore I consider Seyla Benhabib's proposal to think of the status of the preconditions of deliberation in terms of a "historically self-conscious universalism" more promising then Apelt's and Habermas' positions. (Benhabib 1990: 339)

Fortunately, though, we do not need to settle this matter here. As discussed in *Subsection 2.3.4*, for Habermas and those who follow the same research program of rational reconstruction, the status of the preconditions of deliberation and other practices such as the law and the constitutional state are important, because they are the basis of normative theorizing. In respect to TPE, this is not the case. Here, normative theorizing builds on an account of normative individualism and justification. In this context, public deliberation is of interest only as a tool for our model individuals to converge on the ideal state of justified social order as a compromise with ownership. To this end, what matters is that the alleged powers of free-floating argumentation can indeed be harnessed by real groups of reasoners.

In this regard, there is some pertinent and favorable evidence that public deliberation can be achieved by real reasoners and that doing so makes a measurable difference on individual reasoning and collective outcomes. This evidence, also pointed to by Habermas (2009: 150), takes the form of studies on discussion in small and medium sized groups under optimal and controlled conditions.[5] What these studies show is that deliberation can change people's minds and norms can be changed in light of good arguments. So in principle public deliberation works, at least in fairly small groups under ideal conditions, typically involving a well-prepared, structured and moderated discussion.

But of course, it can also go wrong in many ways. (Bicchieri 2017: 158; Parkinson 2006; Müller 2019: 7.2) Especially if we start to think of deliberation as a societal practice. This raises the question of how public deliberation can be implemented and institutionalized so that entire societies can harness the advantages associated with the ideal speech situation. Such matters are typically discussed under the label of "deliberative democracy". Therefore, we now turn to this enormous and still rapidly expending field of research and theoretical debate. But instead of engaging in the hopeless task of trying to give a comprehensive overview on deliberative democracy, my intention here is to focus on two broad strands: *mini-publics* and what is now called *deliberative systems*.

The Macro Perspective

We begin with the macro perspective of deliberative systems. Here the basic idea is that "deliberation can be, and often is, a distributed feature of democracies – indeed, that deliberative democracies [...] necessarily feature a division of labor in which different democratic goods and capacities are activated by different institutions." (Parkinson 2018: 3) Jane Mansbridge and a range of other authors have advocate a "systemic approach" to deliberative democracy, conceiving of public deliberation as something that is achieved by the complex political-societal

[5] In detail Habermas cites James Fishkin (2005; 1995) and Michael Neblo (2010). As we have seen in *Subsection 4.1.3*, Cristina Bicchieri also provides some evidence that group deliberation can be highly effective given certain favorable conditions. (Bicchieri 2017: 156 ff.)

system as a whole. (Mansbridge et al. 2012) This initiative should, however, not obscure the fact that the systemic perspective has been a core aspect of Jürgen Habermas' theory at least since his *Faktizität und Geltung* (1992).

Be that as it may, the perspective of deliberative systems attends to the important question of how a complex organization (the EU, a nation state, a university) can be deliberative. On this macro level André Bächtiger and John Parkinson identify a deliberative, a networked, and a sequenced model,

> "each with its own strengths and weaknesses, and each of which implies somewhat different empirical cues. However, none has been developed to the point where all those cues have been set out and debated, let alone settled. Indeed, each includes significant silences, or over-hasty assurances which fail to fill the silences in any persuasive fashion."
>
> (Bächtiger and Parkinson 2019: 82)

From the systems perspective, the field remains a patchwork of practically underdeveloped accounts because most of the "empirical studies [so far] addressed discrete instances of deliberation, investigated with little if any attention to their relationship to the system as a whole." (Mansbridge et al. 2012: 25) I cannot contribute anything to resolving this gap in research here. Instead, in the following I point to five productive intersections between the perspectives of deliberative systems and TPE.

Firstly, there is the insight that public deliberation is not restricted to institutionalized politics and the law.

> "By contrast, our understanding of deliberative systems includes both informal decisions by accretion and binding decisions that take place outside the state. It goes beyond the boundaries of the nation state to include international, transnational, and supranational institutions, and extends as well to societal and institutional (e.g. corporate) decisions that do not involve the state."
>
> (Mansbridge et al. 2012: 9)

This is a desirable feature of a general framework for thinking about the realization of PE because PE is about justified social norms that are neither confined to national borders, nor to formal political decision-making. A pertinent example here are norms of political correctness that go beyond requirements of the law. These norms can have a strong impact on what people say and what the abstract principle of free speech translates to in everyday situations. And

of course, such norms tend to induce intense public debate, which we sure hope should also aspire to the ideal mode of public deliberation.

Secondly, there is the insight that much about politics that is valuable is not directly linked to the idea of deliberation. From the standpoint of TPE, one important example would be citizens who use civil disobedience to signal that there is something they consider outright unacceptable. Another example is the need for a decision mechanism that produces stable norms in spite of continuous disagreement. We will return to this point below. The lesson here is that deliberative democracy needs to move beyond

> "[...] the idea that deliberative (or any other theory of) democracy captures all relevant politically valuable aspects of democratic practice – democratic deliberation can be understood as one amongst many practices through which democratic institutions and systems realize a range of democratic goods. It is not the only democratic practice and will not always be appropriate."[6]
>
> (Owen and G. Smith 2015: 231)

Crucially, deliberative democracy as well as PE do not offer complete perspectives on politics broadly understood as the discussion and selection of social norms. There are and will continue to be things besides justifiedness or deliberation that matter in politics.

Thirdly, there is the presumption that no matter how much progress the efforts of further specifying the deliberative systems perspective will make, they will not lead to a specification of the one appropriate political system. The best that we can hope for is that in the end we will have fairly concrete criteria that allow us to evaluate the deliberative quality of a given political system. But, even if theorists and scientists converge on a plausible catalog of such criteria, an *eligible set* of actual or possible political systems that could live up to these criteria, would still remain.

Fourthly, there is the hypothesis put forward by Cillian McBride (2015) and Cristina Lafont (2020) that participation in public deliberation itself may cause ownership attachments and avoid political alienation. This points to two research questions at the intersection of TPE and deliberative democracy: One, how and under what condition does public deliberation lead to ownership? Lafont seems to think that it is the mere act of mutual justification that does the trick, but unfortunately she does not offer any evidence that could support her claim or enlighten us about the details. Two, is public deliberation a reliable tool for producing

[6] And while John Parkinson (2018) is very explicit about the fact that the deliberative systems perspective shows that a lot of non-deliberative elements matter, it may well be that not everybody is on board with this sobering conclusion: "From the beginning deliberative theory has had the ambition to provide a normative and empirical account of the democratic process as a whole." (Mansbridge et al. 2012: 24)

bottom-up, group identities based on well-considered reasons? Based on the "Australia Citizen's Parliament" mini-public, Luisa Batalha et al. (2019) argue that group identity can be fostered by public deliberation. According to them, this is done successfully by facilitating a common group identity *and* distinct subgroup identities in order to secure space for critical reflection. These are interesting results. Overall, however, research on the psychology of public deliberation is still in its infancy and, specifically in regard to deliberative systems, it is almost nonexistent.[7]

Fifthly and lastly, a normative background theory such as TPE may prove helpful in providing criteria that allow us to evaluate the deliberative quality of a given system. Jane Mansbridge et al. (2012) offer three allegedly non-controversial "functions" for evaluating the deliberative performance of institutions:

> "The epistemic function of a deliberative system is to produce preferences, opinions, and decisions that are appropriately informed by facts and logic and are the outcome of substantive and meaningful consideration of relevant reasons. [...]
>
> In addition to the epistemic reasons for listening to what others have to say, there are also ethical reasons. A primary ethical function of the system is to promote mutual respect among citizens. [...]
>
> A final function of deliberation, not completely separable from the first two, is to promote an inclusive political process on terms of equality. We call this the democratic function. The inclusion of multiple and plural voices, interests, concerns, and claims on the basis of feasible equality is not simply an ethic added to democratic deliberation; it is the central element of what makes deliberative democratic processes democratic."
>
> (Mansbridge et al. 2012: 11-12)

Of course, Mansbridge et al. (2012) are correct to first stress the epistemic function of public deliberation. This is indeed one of the most uncontroversial aspirations over all accounts in the field of deliberative democracy. It also coheres with the requirement for well-considered preferences in TPE. Nevertheless, in this "most general articulation", it is not very helpful in evaluating any actual system. Neither are the other two items on the list. This is quite understandable, given that the authors are trying to come up with an uncontroversial list that is not committed to a particular normative background theory.

Furthermore, I agree with André Bächtiger (2019: 106) that the compulsive connection between deliberation and democracy is unfortunate. Firstly, because 'democracy' is a highly contested

[7] An exception is the work of Shawn Rosenberg. (Rosenberg 2014)

notion and the site of colliding empirical and normative approaches, whereas the basic notion of deliberation as expressed by the ideal speech situation is quite clear. Secondly, both things are not necessarily connected. There may be democracy without deliberation and there can be (and has been)[8] deliberation in non-democracies.
Fortunately, within the confines of this inquiry we are already committed to one particular normative background theory of justification – namely TPE. Thus we can look at the idea of deliberative politics and polities, independent of what 'democracy' might require. Accordingly, we can attempt a more forthright and detailed list of core deliberative functions from the perspective of TPE and the task of specifying the PEM:

1. **Deliberators:** As always, the individual comes first. If we are to live up to the epistemic and critical tasks of well-considered preferences and critical reflection at all, individuals will have to be endowed with basic deliberative capacities and skills. (Rosenberg 2014) In this regard David Own and Graham Smith (2015: 228) rightly criticize Mansbridge et al. (2012) for not emphasizing that individual citizens should be cultivating a "deliberative stance". On my interpretation this points to the fact that the individuals need an education that teaches deliberative communication and they need the social security (e.g. rights, money and insurance) to practice it. In short, a deliberative system needs capable deliberators.

2. **Deliberative Culture:** That citizens are taught to deliberate and actually do this is unrealistic unless there is a wider culture of deliberation. Here, Owen and G. Smith (2015: 228) make the further important point that the deliberative systems perspective runs the risk of producing "deliberative Schumpeterianism": A scenario in which there are lots of deliberative components but no deliberating citizens. Thus, deliberative quality depends on the existence of an overall culture of valuing and practicing public deliberation, which may be reflected in different forums where citizens actually can and do deliberate.

3. **Diversity:** Even if we have deliberators and actual deliberation, there is still the need to ensure critical reflection and avoid discussions that are flawed by the typical defects of human individual and group reasoning. One effective solution seems to be diversity in perspectives and diversity in identities. Diversity in perspectives helps to avoid exclusion of perspectives and thus the choice of norms that are outright unacceptable to the excluded. It also helps to avoid conservative or majority biases. Diversity in identities helps to prevent a repressive, top-down group identity. (Batalha et al. 2019) This way, it could be crucial in balancing individuality and group identity.

[8]For some examples see James Fishkin et al. (2010) as well as Baogang He and Hendrik Wagenaar.

4. **Deliberative Decision-Making:** If points 1 to 4 are more or less fulfilled and public deliberation is indeed taking place in a systemic way, this should generate knowledge about what kind of reforms of the status quo would produce rejection, acceptance or endorsement amongst deliberators. If, further, the mechanism of political decision- making works on the basis of this knowledge, it should be able to select publicly acceptable or even endorsable equilibrium norms. This is not to say that the mechanism of political decision-making itself needs to be deliberative. But it has to be responsive to the deliberative forces of the overall system in order to reliably select from the set of justified norms.

5. **Inclusion:** Mansbridge et al. (2012) mention inclusion in combination with "equality" and "democracy" – two concepts that have never been very helpful in clarifying anything. My thinking here is that, if we do indeed get to a position where we have a sense of what norms are possible for the right reasons, this would also allow us to focus on potentials for reaching higher levels of justification by inclusion of previously excluded personal values and goals. Especially in respect to those who are troubled by an outright unacceptable norm (e.g. individuals targeted by a racist or sexist norm). Or in respect to those who could be included by simply adding to the list of communally recognized values and goals (e.g. homosexual partnership) or extending the scope of already recognized values (e.g. to disabled persons or non-citizens).

6. **Meta-Deliberation:** "Meta-deliberation is the reflexive capacity of those in the deliberative system to contemplate the way that system is itself organized, and if necessary change its structure." (Dryzek and Stevenson 2011: 1867) What John Dryzek and Hayley Stevenson point to here is of particular importance on my account because it relates to the issue of self-justification. Essentially this should translate into a system's ability to reflect upon anything, even the appropriate conditions and implications of public deliberation, while retaining a robust order.

Now, of course even this extended list is just a preliminary suggestion. Just as with all five points of intersection between TPE and deliberative democracy, the main goal is to show that there is a lot of potential for a productive relationship here. More specifically, TPE can help to focus and guide research and theorizing in deliberative democracy, whereas deliberative democracy can help to relate TPE to social reality and thus help to specify what it actually requires of citizens and their institutions.

The Micro Perspective

Let us now turn to the micro perspective of mini-publics. Mini-publics "are independent and facilitated group discussions among a (near) random sample of citizens who take evidence from experts and interested parties." (G. Smith 2018: 4) Usually these group discussions between 12 to 200 and more participants are carefully planned and moderated events that come in several different formats ("citizens' jury", "citizens' assembly", "deliberative poll" etc.). The goal is to come up with proposals on some predetermined political topic. Essentially this is the closest we have to the idea of an ideal speech situation or the forum of public deliberation in real politics. The institutional role of this tool ranges from being a mere experiment or a forum of civic participation, all the way to being a constitutive part of political decision-making.

In contrast to the deliberative systems and focusing on the macro level, mini-publics have three distinctive advantages. One, studies of small to medium sized group deliberation provide most of our empirical evidence that and how public deliberation works. Doing similar research on the scale of entire systems would be challenging.

Two, mini-publics are a familiar and effective way of making regular citizens part of the abstract ideal we keep talking about. Thus it is a way of fostering a culture of being capable and actually participating in public deliberation. Just imagine every community (city, village, municipality) having regular mini-publics on important decisions that actually feed into political decision-making. This probably would check a lot of boxes on any plausible list for evaluating systemic, deliberative quality.[9]

Three, mini-publics can be incorporated where institutional innovation is needed most: political decision-making. With mini-publics we do have a means to infuse the powers of the ideal speech situation into political decision-making right now. This is an urgent matter because as it stands our democracies suffer a bias toward capitalist and overall shortsighted human interests. (MacKenzie 2016) This defect is not merely a worry of not being close enough to utopia. On the contrary, there is a very real possibility that this defect and the associated incapability to switch to a sustainable economy (fast enough) leads to a scenario where substantial amounts of humanity will perish.

None of this is to say that mini-publics are a cure for everything. However, I am explicitly pointing to these advantages because the narrative of Mansbridge et al. (2012) suggests (without explicitly claiming!) that mini-publics are the main idea of a past stage of theorizing and

[9] "Deliberative democratic theory is full of statements about the general facilitating conditions – in particular, the rights, principles and dispositions – necessary for the emergence and sustenance of public deliberation between free and equal citizens. However, our analysis of mini-publics, PB and internet discussion forums, in particular, highlights the fundamental role that active facilitation plays in realising such rights, principles and dispositions. Citizens do not necessarily come fully formed in a deliberative sense: facilitators continually shape and reshape the conditions for deliberation." (G. Smith 2009: 197-198)

research in the field of deliberative democracy, whereas deliberative systems are the preceding, current game in town.

Mansbridge et al. (2012) and Graham Smith (2018) are certainly correct to point out the limits of mini-publics. These limits consist, first, in the fact that they are not self-sufficient, but usually require organizers, hosts, moderators, agenda setters and so forth. Second, how the discussion and outcomes of mini-publics are communicated to the larger public and transformed into binding rules is also a difficult issue. Thus it does indeed make sense to think of mini-publics and deliberative systems as complementary approaches.[10]

Nevertheless, I feel the need to emphasize that it is difficult to predict if and when the systemic approach will lead to paradigmatic framework and practically relevant outcomes. At the same time we already know that and how we can bring instruments such as mini-publics to bear on pressing problems in democratic decision-making. Having such a tool is valuable, especially because – as we will learn from Jeremy Waldron shortly – in politics there is usually a deadline. And especially in light of a shrinking habitat for human beings on earth, even normative theorists will have to increasingly take into account such practical urgencies.

This concludes our brief excursion into the field of deliberative democracy. As we have seen, this extensive field of theorizing and research allows us to think of the abstract ideal of having, if you will, a *forum* of public deliberation in a somewhat more practical manner. More precisely, deliberative democracy offers a way to think about how the forum can be translated into social reality at different levels. Here, I have focused on the macro level of deliberative systems and the micro level of mini-publics and emphasized that both should be given equal weight. Further, I hope to have shown that deliberative democracy is a suitable background condition for the core tasks of the PEM to be fulfilled. That is, before the background of deliberative democracy it becomes more tangible that citizens have well-considered preferences on matters of social order, critically reflect on the status quo and reforms, flag existing or proposed norms they consider outright unacceptable, and gain a sense of their eligible set of reforms and what alternatives might lead to a more justified social order.

However, we have also come up against several impasses. One such hindrance is the ragged state of theorizing in deliberative democracy, especially when considered the systemic per-

[10] "This helps make some sense of an ongoing accusation that some partisans of minipublics press against the systems approach in particular. While it is certainly the case that the deliberative systems approach is partly a reaction against thinking that deliberative minipublics – and other micro arenas and forums – are 'deliberative democracy', sufficient unto themselves, pointing out the limitations of minipublics does not mean that one condemns them as entirely useless; on the contrary, it could be a highly constructive move [...]. What follows is that one does not have to choose between micro and macro visions as if they were competing orientations to the same phenomenon; instead they can be seen as complementary, nested orientations which address different aspects of something bigger." (Bächtiger and Parkinson 2019: 107)

spective. Another problem is the empirical uncertainty as to whether and how deliberation leads to ownership. Finally, the discussions in this subsection have brought out important considerations that go beyond the forum of public deliberation, such as the preconditions of deliberation and the need for a decision-making mechanism. This points us to the need for a wider kind of proceduralism, which takes into account that establishing justified social order in a diverse society is about more than public deliberation. To this end, we will consider some insides from Jeremy Waldron's democratic proceduralism and, later on, Gerald Gaus' efforts in modeling the open society.

5.1.3 Jeremy Waldron's Democratic Proceduralism

Jeremy Waldron's "democratic proceduralism" jointly addresses aspects of public deliberation, political decision-making and basic rights under conditions of reasonable and deep pluralism.[11] Waldron's argument proceeds from a fundamental concern for individual autonomy and a universal right to justification. (Waldron 1987) Although I started out with a normatively more modest conception in *Chapter 1*, our accounts converge on many points. Perhaps because Waldron focuses on pluralism and disagreement with a firm focus on actual institutions and thus proves resilient against the temptation of searching for substantial normative principles of morality or justice:

> "I believe that philosophers of public affairs should spend less time with theorists of justice, and more time in the company of theorist of authority and theorists of democracy, reflecting on the purposes for which, and the procedures by which, communities settle on a single set of institutions even in the face of disagreement about so much that we rightly regard as so important."
>
> (Waldron 1999: 3)

I could not agree more, although I would add that the "philosophers of public affairs" should also spend much more time with scientist in order to sharpen their empirical models. In the following I summarize and discuss several of Wadlron's insights, as they specify and widen our understanding of the PEM.

The first two points relate back to our discussion of public deliberation above.[12] More specifically, Waldron clarifies, one, that it is a mistake to focus on an ideal of consensus and, two, that public deliberation ought not to be restricted by a requirement of public reason.

[11] I have benefited greatly from an overview article on Waldron's work by Fabian Wenner (2013). Referring to Waldon's theory as "democratic proceduralism" is suggested by the title of Werner's paper.

[12] Waldron (1999) uses the term 'deliberation' mainly to refer to discussions of elected representatives in parliament.

The Persistence of Disagreement

The first point originates with Waldron's rejection of Joshua Cohen's idea of "deliberative democracy", according to which the aim of public deliberation is "to arrive at a rationally motivated consensus". (J. Cohen 1989: 23) In this context, Waldron also endorses consensus as the appropriate ideal that guides the logic of deliberating citizens who should try to convince each other with arguments that are acceptable to all. Nevertheless, Waldron maintains that the normal outcome of any deliberation is still reasonable disagreement, which, perhaps, is more reasonable after the exchange of reasons than before. The broader conclusion Waldron draws from this point is that democratic proceduralism under the circumstance of politics equally requires public deliberation *and* a mechanism of political decision-making such as voting. (Waldron 1999: 91-93)

Although I agree with these conclusions, I am not sure whether there is any important disagreement between Waldron and Cohen.[13] Be that as it may, I am not aware of anyone who thinks that ideas of public deliberation and deliberative democracy are about eventually resolving all debate and reasonable disagreement in light of unique, rational outcomes. Nevertheless, it is worth emphasizing the illusive character of 'consensus' in this context because some continue to be put off by the confused impression that public deliberation and deliberative democracy are all about consensus.[14]

Unrestricted Deliberation and Justification

Waldron's second invaluable point for getting at an appropriate conception of public deliberation is his insistence on the openness of the discussion to the range of reason. The heart of his argument is this:

> "[The] idea of justification in itself involves no restriction on the range of reasons that it is appropriate to mention. [...]. So, justification is open and inclusive. It is interested in any reason there might be for or against [some decision] D, and reasons can come from unexpected directions."
>
> (Waldron 1999: 116-117)

[13] The difference seems to be in the contrast between the idea of a "rationally motivated consensus" (Cohen) and a state of enduring disagreement (Waldron). Note, however, that this differentiation only holds if 'consensus' implies that every party gets their most preferred outcome and thus a consensus can never be the kind of pragmatic settlement (what I call a "compromise") that Waldron and I are arguing for. If, however, a consensus can also be a compromise that, as Cohen puts it right after the quoted passage, is "persuasive to all", e.g. in light of the personal and pragmatic reasons they have, then the two positions collapse into each other. Thus the difference between Waldron's democratic proceduralism and Cohen's deliberative democracy only holds if Cohen believes that the point of deliberation is to terminate disagreement by establishing to universal agreement *on the best outcome*. The difference evaporates, however, if the deliberative "consensus" is merely about a rationally persuasive outcome for all.

[14] Consider for instance the hasty rejection of deliberative democracy by Gerald Gaus (2011: 387).

> "[W]ithout this requirement of openness, the reasoning process that justificatory deliberation involves is in danger of becoming not just truncated but distorted. We will not be in a position to determine the true weight or bearing of the reasons that we consider unless we take into account the weight and bearing of all the reasons that are in fact relevant to the weight and bearing of the reasons we consider."
>
> (Waldron 1999: 121)

Further, any restriction to shared reasoned is not necessary for actual public deliberation because it is the process itself that is meant to provide for the necessary filtering:

> "Comprehensive ethical, philosophical, or religious doctrines are not excluded from public reason because they are wrong or false or ideological. No doubt some are. But we do not need any special doctrine of public reason to justify the exclusion of reasons resting on false beliefs or false or invalid moral or ethical principles or reasons that have no real relevance to the decision in question. Basic rationality copes with that."
>
> (Waldron 2007: 109)

Waldron's insistence on the openness of deliberation is primarily aimed at John Rawls' conception of public reasoning as reasoning on shared (thus "public") instead of non-shared reasons that are part of people's comprehensive and conflicting world views. (Rawls 1997: 800) Besides pointing to the problem that Rawls' conception forces people to neglect the reasons which are perhaps most relevant to them in a give context (e.g. religion-based reasons in a debate about abortion), Waldron goes on to show that public reasoning is actually neither a fitting nor a desirable description of Rawls' favorite example of real-world public reasoning: the legal rulings of judges.

Waldron is absolutely right to insist that justification is open in the sense that it does not require any restriction to shared reasons. Consequently, also public deliberation – the practical realization of the justification requirement – is also open in this way. It is only restrained by how the notion of having good reasons as specified in the justificatory account and what is involved in reasoning as a practice. In this respect it is quite telling that in the field of deliberative democracy demands of civility in deliberative engagement, are more common than requirements of shared reasons. (March and Steinmetz 2018)

I emphasize this point here because the issue continues to be debated in public reason theories, which start from a requirement of public justification. That is, a requirement "that the moral or political rules that regulate our common life be, in some sense, justifiable or acceptable to

all those persons over whom the rules purport to have authority." (Quong 2018). In respect to this requirement, there is an ongoing debate as to whether appropriate reasons for meeting this requirement of justification need to be shared by all, accessible to all, or merely mutually intelligible. (Vallier 2011; Quong 2018)

This is a confused debate. Perhaps the confusion is also due to the influence of hypothetical choice modeling. As we have seen in *Chapter 2* this strategy typically involves the theoretical uncovering of universally shared reasons and does not handle diversity and non-shared reasons well.[15] Be that as it may, as Waldron points out, the idea of justification does not include but rather excludes any shared reason requirement. If some individual i has her own good (well-reflected, undefeated etc.) reasons in favor of social state S, S is justified to i. To give an example, if an economist and an evolutionary biologist both confirm to the same account of social norms, they are both justified in endorsing that account, even though they may have quite different kinds of theories and evidence – i.e. different reasons – in favor of it and we might as well also assume that these reasons are mutually unintelligible. Note also that having the same reasons (e.g. because now both are economists) would not make any difference to how justified each one is in endorsing the account in question. Justification is about having good reasons, not shared reasons. The respective difficulty consists in specifying what "good reasons" are.

The practical consideration that is often confused with the matter of justification is the desirability of shared reasons for the stability and efficacy of a cooperative order. As discussed in *Subsection 4.2.3* shared reasons and understanding are indeed helpful for facilitating trust, symmetric expectations, stable norms, group identity, conflict resolution and reform. Stephen Macedo (2010) has elaborated on several of these advantages and he is absolutely right to present shared reasons and shared understandings as a desirable aspiration for cooperative social order. Unfortunately, he also presents his argument as a contribution to the confused *consensus vs. convergence debate* in public reason theory. Thus adding to the confusion that the idea of justification and the desirability of shared reason for shared norms are related.

Again, justification does not require, but is hindered by a requirement of shared reasons, and public deliberation is just as much about the reasons we don't share (yet) as it is about what is implied by the reasons we already share. Therefore, the interesting practical question of shared reasons should not be asked on the level of normative theory, but on the level of social science and psychology, where interesting answers are likely. Getting back to the conceptual point of the very beginning of this section: from a practical perspective, public justification is

[15] "[T]he dominant public reason views have been committed to strong normalization: public reason has typically been identified with the reason of the normalized public perspective – the liberal perspective." "[S]upposing that we approach political philosophy through a normalized, or common, perspective on justice." (G. Gaus 2016: 168; 145)

about a justification that is given to all, in public, but not necessarily shared by all.

The Right Mechanism of Decision-Making

This brings us to the points raised by Waldron that go beyond the idea of public deliberation. The first important point here is that, since public deliberation does not produce consensus, we need a political decision-making mechanism that selects one possible norm over another, even if there remains reasonable disagreement on the matter. This point in turn is based mainly on what Waldron calls the "circumstance of politics". As we already know from the discussion in *Subsection 4.2.1*, this denotes the assumption that politics is about having to coordinate on one of several possible regulations while citizens continue to reasonably disagree on the best or correct regulation. Further, in politics there is deadline: At some point we need a decision and move on in spite of persistent disagreement. (Waldron 1993: 34-35, 1999: 101 ff.)

With this in mind, the obvious subsequent question is about the appropriate mechanism of political decision-making. Here Waldron's focus is on voting and the majority principle because he claims that this principle best reflects the respect for the individuals under the circumstances of politics. More specifically, he argues that voting expresses the equal respect of persons and accounts for the fact that a majority actually favors something. (Waldron 1999: 111-114) He further argues against a system of judicial review of constitutional courts as a counterweight to voting. (Waldron 1999: 285 ff.)

I, for one, am not so enthusiastic about majority voting. It is not that there is anything wrong with this mechanism but rather that I believe that we should be open to consider many different mechanisms and the different advantages and disadvantages they imply. So in respect to the problem of selecting from the eligible set of of justified norms, I think we firstly should acknowledge that reasonable disagreement certainly extends to the question of the (most) appropriate mechanism. (G. Gaus 2011: 391) Waldron argues that majority voting best satisfies the basic liberal values that are fundamental to his theory. However, these values are not fundamentals of TPE and even if they were, I still would object that majority voting does not *uniquely* satisfies these values.[16]

Secondly, I believe the selection of the most appropriate decision mechanism mainly turns on complex practical consideration that we cannot settle in theory. One of these considerations is decision-making costs. According to James Buchanan and Gordon Tullock (1962), decision-making costs depend on whether the decision-making procedure is rather inclusive or exclusive. As more people are included in the procedure, the costs of decision-making rise, while external

[16]From a more practical perspective one may further object that in our representative democracies, a particular voting mechanism in parliament or in the election of representatives often does not lead to policies that are preferred by any majority of citizens. Neither are these voting regimes very effective in securing political equality.

enforcement costs fall. Conversely if less people are included in the procedure, dicision-making costs fall, but external enforcement costs rise.[17] Besides costs, stability is another obvious consideration. As already pointed out in *Subsection 3.2.3*, any robust social order implies some compromise between being able to change the norms and the need for having stable norms. Further, if we think of robustness in more systemic terms, we might be motivated to establish a system of several different decision-making mechanisms.[18] Not only because different mechanisms may be more or less appropriate in different contexts, but also because having different mechanisms can establish a system of checks and balances. For example, the legislative may propose a law that is made subject to a popular vote and additionally revised by judicial review. Thus, essentially, due to complex practical matters involved, we cannot recommend *the* appropriate decision-making mechanism, or combinations thereof, in theory. This point also extends to Waldron's discussion of rights, which we will now turn to.

Securing Individuality Through Rights

Besides the need for a decision-mechanism, Waldron also points to the importance of basic rights – in particular political rights of participation. The precise content of these rights is a source of persistent disagreement. Thus we should not think of them as inalienable rights that are beyond disagreement and adaptation. In particular, Waldron argues against these rights being enshrined in a bill of rights. (Waldron 1999: III.10)

This point relates to a more comprehensive quarrel with Ronald Dworkin about whether the American system of a written constitution, a bill of rights and a constitutional court engaging in judicial review is overall preferable to the British system that does without any of these things. Again, as in the case of the appropriate procedure of decision-making, Waldron's initial analysis is convincing, but he ends up taking it too far in that he prescribes specific institutional regimes.

More precisely, Waldron is absolutely correct in pointing to the need for a decision-making mechanism beyond deliberation as well as individual rights and political freedoms. These rights produce a secure space for the individual to gain and voice a critical stance on existing and proposed norms. And there is a whole range of rights and freedoms (freedom of speech, freedom of the press, freedom of movement, freedom of association) that are obvious practical

[17] Also, from this perspective there is nothing special about majority rule: "[I]n our preliminary analysis, once the rule of unanimity is departed from, there seems to be nothing to distinguish sharply any one rule from any other. [...] Moreover, on a priori grounds there is nothing in the analysis that points to any uniqueness in the rule that requires a simple majority to be decisive. The (N/2 + 1) point seems, a priori, to represent nothing more than one among the many possible rules, and it would seem very improbable that this rule should be "ideally" chosen for more than a very limited set of collective activities. On balance, 51 per cent of the voting population would not seem to be much preferable to 49 per cent." (Buchanan and Tullock 1962: 64)

[18] "[I]t will be rational for the individual to choose more than one decision-making rule for collective choice-making under normal circumstances." (Buchanan and Tullock 1962: 63)

preconditions for the forum of public deliberation to emerge at all. The detailed formulation and application of such rights will of course remain controversial.

Nevertheless, non of this settles the trade-offs involved in choosing between political (constitutional) system. To exemplify, I do not see why inalienable rights or a bill of rights should be out of the question. Having a bill of rights does not mean that these rights are beyond debate and change. Effectively, the meaning of these rights is continuously changed by new legislation and court rulings that specify what these rights imply in a given situation. And this way of reinterpretation may serve a society just fine until they feel the need to produce a completely new list of rights in the form of some procedure of constitutional renewal. Sure, on the continuum between stability and flexibility, constitutional entrenchment of rights is quite far on the side of stability. Nevertheless, they are still just constitutional norms and not holy commandments – i.e. they are not beyond critical debate and alteration. And if a society considers this level of entrenchment worth the price, for instance because it is thought to be an effective protection against some of the worst monstrosities human beings can engage in, I do not see a cogent objection in Waldron's work.

My general conclusion here is that, and I am sure Waldron would agree, any regime of rights or decision-making mechanisms should be a potential topic of public deliberation. However, unlike Waldron, I suggest abstaining from prescribing specific institutional regimes because, on this level of abstraction, they do not follow. Settling on a regime of rights and political decision-making is a joined effort of theorists, scientists, other experts, politicians and, last but not least, a public of citizens.

Circular Justification

Let us move on to another tricky point, also raised by Waldron: If it is disagreement all the way down (up, left, or right) and thus everything is "up for grabs", how can there be any stability? (Waldron 1999: 303 ff.) Or, put slightly differently, how can proceduralism solve any problems if the procedures themselves are controversial? (Raz 1999: 47)

I think the core lesson from reasonable disagreement on all levels and the requirement of self-justification is that we should try to come up with a robust mechanism while resisting the temptation of trying to make it immune to fundamental challenges and rejection. Waldron's interpretation of this lesson is that we should not withdraw social order from majority voting by entrenching it in a constitution that is difficult to alter. I have tried to show that these are practical considerations that can be resolved in different ways. To me it is sufficient that we refrain from putting in place linguistic or other restrictions that inhibit the critique or rejection of deliberative principles in the forum of public deliberation. This means that we do not eliminate the possibility that, for instance, someone openly speaks in favor of exclusion,

racism and forceful domination of some minority. For if we do not allow for certain thoughts to be thought or opinions to be expressed, we are only producing another kind of indoctrination that does not respect the basic idea behind justification: individuals having their own good reasons for some social arrangement. The consequence of this openness and potential for self-rejection is that public deliberation, liberal order and democracy may be destroyed from within. This is not a small price to pay. But allowing for this possibility is the only way to stay coherent and ensure that, if these things endure, they do so for the right reasons.

But of course, no order should actively seek its own rejection. Quite to the contrary, we further need to explain how it can be possible to have a robust social order that successfully fulfills its function of coordinating individuals on cooperative behaviors, while these very individuals keep disagreeing on the how and what of coordination. One straightforward solution is what we usually call "reforming". This means that we only discuss and change social order one piece at a time, while leaving other aspects untouched and avoiding anarchy and chaos by only practicing successive replacement. Now of course, every reform needs a procedure. And all procedures are potentially controversial. But if a group has more than one procedure at its disposal for settling different aspects of social order, it could always use one procedure to change or replace the other. In any event, what is always needed is the ability of a group or society to settle, for the time being, on one procedure to proceed with. And while there sure does exist the possibility of getting lost in disagreement here, for the most part human societies seem to handle this problem just fine.[19]

In conclusion, we do not get anywhere without accepting some procedure in face of reasonable disagreement. But I do not see any fundamental problem here, for human beings seem to be quite capable of coming up with workable procedures, and as they move along, there is nothing keeping them from distinguishing the more from the less reasonable procedures.

5.1.4 Gaus' Recommendations for an Open Society

At this point I briefly return to Gerald Gaus and his (2016) *The Tyranny of The Ideal*. As already mentioned at the very end of *Chapter 2*, in this later work Gaus also argues for an open-ended ideal he calls, following Karl Popper, "the open society". Gaus' "open society" is in many ways about considering what a "well-ordered society" could look like if we take diversity and fundamental disagreement on "justice", or any other ideal, seriously. In this respect, his and my take on what normative social theory can hope to achieve convergence.

[19] If I were pressed to speculate about the explanation of this phenomenon, my best guess would be that citizens accept a given procedure because they have a "presumption for reasonable outcomes" (Habermas 2009: 413). That is, on my account, they believe that the procedure is effective in selecting options from the realm of mutual benefit. This presumption may be based on experience or familiarity with the procedures of one's cultural.

The main difference remains that Gaus is not interested in the idea of public deliberation. Instead he focuses on ever more refined models of the problem of diversity and reform. He further has to offer "some tentative conclusions about the sorts of institutional structures and principles that are friendly to diversity per se." (G. Gaus 2016: 176)

The Principle of Natural Liberty

The first suggestion along these lines is Gaus' argument for a principle of "natural liberty" as a closure principle:

> "The public moral constitution of the Open Society, then, is largely a morality of prohibitions and requirements, for such a morality allows individuals maximal opportunity to explore novelty and diversity, and so explore their perspectives while still possessing a shared moral constitution — a common public world — via which they can coordinate their activities and advance claims against each other employing public rules and categories."
>
> (G. Gaus 2016: 198)

This conclusion is based on a fairly complex argument, which starts from the claim that any set of norms practically requires a closure principle, telling people what to do in unregulated cases. Gaus further distinguishes two ways of fulfilling this requirement. One way is to have a set of norms in the form of prohibitions and "natural liberty" as a closure principle: whatever is not prohibited is permitted. Another way is to have a set of permissive norms and "residual prohibition" as a closure principle: Whatever is not permitted is prohibited. The last and crucial step in the argument is that the first variant, the reign of natural liberty, is more beneficial, in that it is an enabler for creativity, innovation and productivity in diverse society. There is something highly plausible in Gaus' argument. And I think this is because in the West we are already convinced that a system that is liberal rather than prohibitive is better in that it is more pleasant and performs better in many respects. Irrespective of this general plausibility, there are also good reasons to be careful here in deriving more substantial conclusions. One such reason emerges as we step down from the abstract systemic perspective and turn to concrete areas of regulation. Consider for instance that you would want different closure principles in the case of innovation in ice cream flavors and in case of innovation in defense technologies. There are simply very different levels of risk involved here. And besides different objective levels of risk, there are different subjective, cultural attitudes toward risk. There is, for instance, the notorious difference between Americans and Europeans in whether new technologies are rather associated with risk or with opportunities and whether such technologies need approval to begin with.

My overall point here is that, on a fairly abstract level, Gaus is right: There seem to be important advantages associated with living in a liberal rather than a prohibitive system. This, however, leaves plenty of room for different closure principles in respect to different areas of regulation and different (cultural) stances on risk-taking. Essentially, the PEM should be open to both kinds of closure principles, while perhaps granting the benefit of the doubt to the principle of natural liberty.

Polycentric Order

Another suggestion for the diverse society promoted by Gaus is "polycentric" social order. This denotes the idea that social order consists of different "social networks" (e.g. vegetarians, religious citizens, feminists, and libertarians) with different internal norms. The advantage of this difference in norms in different subgroups is, firstly, that it allows for diverse ways of living and thereby reduces the need for collective decision-making. Secondly, it allows for a competition of different norms which might eventually spread to multiple subgroups or the entire population. In other words, this system allows for existing diversity to be lived and to be of potential benefit to everybody.

Gaus makes this argument in favor of polycentric order only in regard to "moral norms" because he adheres to, especially in light of his own work, a more and more obscuring distinction between the moral and the political.[20]

However, there is also a more political variant of polycentric order put forward by Julian Müller (2019), who argues for a "polycentric democracy" on the same basic motive of turning diversity into an asset rather than a problem. Müller makes a range of interesting points. Of particular relevance in this subsection is his claim that group discussions tend to have a "conservative bias" in that they tend to "disregarding novel ideas and solutions, while emphasizing beliefs, institutions and values that are already established." (Müller 2019: 106) He defends this claim, on the one hand, by reference to a range of psychological studies on biases and defects in group discussions[21]. On the other hand, he reference to the costs, efforts, limited individual capacities, complexities and uncertainties that practically limit what can be achieved by group

[20] On this, see Gaus' discussion of *The Moral and Political Constitutions*, where he shows by reference to the work of Garry Mackie and Marion Young how the law may fail to govern "social norms", while failing to appreciate that this is also a great example for why law and morality in action are the same thing. (G. Gaus 2016: 206-207)

[21] The cited works by Müller include: Sunstein, C.R. (2006): *Deliberating Groups versus Prediction Markets (or Hayek's Challenge to Habermas)*, Episteme, Vol. 3 No. 03, pp. 192–213; Stasser, G. and Titus, W. (1997): *Effects of Information Load and Percentage of Shared Information on the Dissemination of Unshared Information During Group Discussion*, Journal of Personality and Social Psychology, Vol. 53 No. 1, pp. 81–93; Asch, S.E. (1951): *Effects of Group Pressure upon the Modification and Distortion of Judgments*, in Guetzkow, H. (Ed.), Groups, Leadership and Men, Pittsburgh, PA, pp. 177–190; Larson, J .R. (2010): *In Search of Synergy in Small Group Performance*, New York, NY.

discussions.
Müller's main conclusion from this analysis is that we should look for an alternative way of dealing with diversity. His suggestion is that we focus on the idea of a structured competition between different polities, inhabited by more homogeneous individuals who converge on a common vision of utopia. In more practical terms, Müller suggests that states should allow for existing or new-founded cities to become "free cities". These are assumed to attract a rather homogeneous citizenry, interested in living under an order that is radically different from neighboring polities and the order of the overarching polity (i.e. the nation state).

Overall, I think Gaus' and Müller's reference to polycentric social order and the potential of diversity is of direct relevance to the PEM. Specifically a subdivision of social order may facilitate orders that divers groups of individuals can endorse and identify with. Müller's critical reflections on the biases often inflicting group discussion are also of relevance for designing a social mechanism around the idea of public deliberation. Unfortunately Müller does not discuss in detail what the cited findings mean for theories of public deliberation and deliberative democracy more generally. Helene Landemore (2013) for instance argues, that there are specific tools we can use in organizing public deliberation, in order to prevent many faults in group reasoning. She also argues together with Hugo Mercier that human reason is more effective when reasoning in groups then reasoning alone. (Mercier and Landemore 2012) Also unfortunate is Müller's reluctance to critically asses his preferred approach of polycentric order.[22] Nevertheless, I agree with Müller on two important points. One, public deliberation is a demanding ideal that does not solve everything and has to come to terms with several common defects of group discussions. Two, in light of this, having space where individuals and subgroups can experiment and compete seems to be highly useful in the quest for justified social order.[23]
I think in regard to polycentrism and the PEM the decisive question is: Where do people think that they have common problems that need to be resolved by common norms? Wherever there is a perceived need for common norms (e.g. an official language), polycentric order is not a helpful answer. But wherever there is room for different and competing orders, it might very well be the right answer. Yet, as I keep pointing out in this subsection, there will be

[22] That is, he does not consider problems that might result from dividing people into different, fairly homogeneous bubbles. This might, for instance, cause them to be less capable of engaging and discussing with others who think differently. At the same time there certainly are pressing problems facing the larger community and perhaps all humankind, which require high degrees of more (global) collective decision-making.
[23] I also endorse the idea of reintroducing "free cities" because, as Müller points out in his criticism of group discussions, there are some political issues and potentials for reform (e.g. think of the idea of unconditional basic income) that are so complex and uncertain in their effects that at some point we need to stop the discussion and start (large-scale) experimenting.

controversy. And this point certainly extends to the perceived need for common norms. Thus the appropriate institutional setup, e.g. the appropriate degree of federalism, of freedom granted to "free cities", or to every individual in her private sphere, remains indeterminate on the level of normative social theory.

Concluding Remarks on Mechanism Design

This concludes our reflections on mechanism design. Admittedly, we have not come very far. What we have gained is a general idea of what a more comprehensive discussion of the PEM would be about: A wide proceduralism consisting of two things. Firstly, a regime of public deliberation ensuring that norms are selected or persist as a function of well-considered individual preferences. Secondly, a range of further mechanisms and considerations including a regime of rights and freedoms, mechanisms of political decision-making, as well as space for diversity and experiments in living and in polity.

A further conclusion of this subsection is that, beyond such considerations of a wide proceduralism, from the perspective of theory we should remain neutral on specific institutional setups. This does not come as a surprise in the context of a discussion that started out with considerations of equilibrium norms and eligible sets. With such a background, many things remain open-ended on the level of abstract theorizing and eventually need to be specified by more local theories, (social) science, politics and citizens of a given society.

This is not to say, however, that we should not pay attention to the many concrete and innovative ideas for institutional design such as incorporating mini-publics into political decision-making or allowing "free cities" to experiment with radically different orders. We should simply stress that these arrangements are not deductively handed down from normative theory to politicians and citizens. Rather, they should be considered as refined suggestions of how certain ideals can be pursued in social reality.

5.2 An Index of Justified Social Order

As we have just seen, the choice of specific institutions constituting the PEM remains inconclusive on the level of theory. Hence, and as already pointed to in *Subsection 3.2.4*, more specific prescriptions do not follow deductively from an open-ended ideal such as justified social order as a compromise with ownership. It is rather up to a joined effort with other academics, experts and citizens to come up with proposals for a specific PEM in a given society and there is no reason why different groups should not come up with different institutional regimes for striving toward the ideal. All the more desirable in light of this inconclusiveness is the prospect of an independent test of how well a given society is in actually realizing PE. The outline for such a

test is the topic of this final section of our inquiry.

The basic idea, as explicated in *Subsection 3.3.2*, is that individuals who participate in a social practice on a voluntary and well-informed basis signal agreement to the rules of said practice. This ideas can be applied to many different contexts. To my mind, the most illustrative case is partaking in a game, such as chess, and thereby signaling agreement to the rules of said game in virtue of starting to play. In a similar way, acts of voluntary participation in politics, such as voting, can be interpreted as signaling agreement to the rules of politics. Even informal norms such as "Bring a gift if invited to a birthday party!" could be said to be consented to by the mere fact that individuals keep gifting when attending birthday parties.

When thinking about these examples and whether they are actually plausible scenarios for signaled agreement, we are challenged to think about three things: First, how can we ensure that the background conditions of voluntary and well-informed choice are actually met? Second, what are the relevant types of participatory acts that signal normatively meaningful agreement for some given instance of social order? Third, how does the observance of relevant acts of participation under favorable conditions of voluntary and well-informed choice combine into one testing conception of justified social order?

These three questions translate into three steps, which we will attend to successively in this section. The first step is a discussion of the essential aspects of systemic public deliberation for epistemic purposes and how they can be observed in practice. The second step is all about identifying participatory acts that signal significant normative agreement according to our normative model. The third step consists in proposing a way of combining the observance of relevant acts of participation and of voluntary and well-informed background conditions into an index of justified social order.

Eventually what I want to end up with is an empirical testing conception – an index – that allows us to score the level of justification of a given social order. In virtue of this measure, different social orders, or the same social order, could be compared over time. The Index could of course not test the justifiedness of merely proposed norms and orders. This is why the idea of having a social mechanism for pursuing Political Equilibrium (PEM) and a testing conception work in tandem: The testing conception can evaluate the status quo, but it will not produce reforms. The PEM, on the other hand, will hopefully produce improving reforms, but it lacks an independent test of how well it is actually performing.

The overall goal of this section is to show that and how a testing conception of justified social order in form of an index could be constructed. One restriction implied by designing an index is the focus on the formal side of social order. That is not to say that informal social order is not also a pertinent object of the testing conception. However, if we focus on informal orders, we should think about a different kind of test than a tool from democracy research. Devising

such a tool is beyond the scope of this inquiry. Thus I confine the following reflection to the quest for an index of justified, *formal* social order.

5.2.1 Systemic Public Deliberation as a Necessary Condition

The first task to be dealt with in order to get to a testing conception of justified social order is to translate the conditions of voluntary and well-informed choice into social reality. We begin with a brief discussion of voluntary choice before turning to well-informed choice, which is the core problem of this subsection.

Being Free to Choose

Voluntary choice is a complicated matter in a wider social context because in society your choice set is usually dependent on the wants of others and given circumstances. So there is no absolute freedom to be had in society. You cannot simply do whatever you want – anywhere. The complicated problem, then, is to settle on the proper amount of freedom we want to grant each other in society.

Fortunately, we do not need to wrestle with this complex political issue here. In our case, the question is merely; what is the right kind of freedom required for a choice to signal normatively significant agreement? In this context, the matter of voluntary choice is fairly trivial, for we simply need an isolated, well understood choice situation that is not distorted by hidden incentives or considerations which should be irrelevant. For instance, if we consider a scenario where somebody is choosing to play chess with a friend, we assume that she is not being offered any money for playing and that her livelihood is not affected by the choice in any way.

Well-Considered Choice through Public Deliberation

It has been a common theme right from the beginning of this inquiry and the formulation of the JPN in *Subsection 1.3.2* that justified social order must rest on well-considered, individual reasoning and preferences. Now, as discussed in *Subsection 1.2.2*, it is highly sensible to think about this as a contextual standard. That is to say, it does make a huge difference whether we are considering a scenario of choosing between going for a bike ride or a swim in contrast to a scenario where we choose between different constitutions. The practical standards of what one might call *sufficiently* well-reflected preferences and choice differ substantially in both scenarios.

Looking at the complex social orders we have in and beyond the nation state, it is safe to say that it will be difficult to live up to the standards of well-considered choice. This is especially true for ordinary citizens who do not earn their living by thinking about these matters in the first place. Therefore, we cannot simply demand or assume that citizens have well-considered

preferences in respect to the question of social order. Rather, as I stressed in the discussion of systemic public deliberation above, we need to ensure that there is a framework in place which assists citizens in getting as close as possible to having sufficiently well-reflected preferences on social order.

This perspective presents a purely epistemic route to the idea public deliberation. Therefore, in the context of the testing conception, we are interested in public deliberation primarily as a way of ensuring well-considered individual choice. Whereas in the previous section, public deliberation was meant to ensure well-considered collective choice of norms (i.e. political decision-making). In this section and in respect to the testing conception of PE, it does not matter how given norms where chosen. They might as well have been chosen by a monkey rolling a dice. What matters is that public deliberation fulfills its epistemic function of producing well-considered individual preferences on social order and thereby rendering individual acts of participation normatively meaningful.

Having these preliminary considerations out of the way, let us turn to the questions of what constitutes a deliberative framework for the facilitation of well-reflected preferences on social order. Now, there are some empirical measures derived from the ideal speech situation, such as the *Discourse Quality Index* and *VisArgue* that are of relevance here.[24] However, the problem with these measures in our context is that they are specifically designed for measuring the deliberative quality of particular acts and venues of deliberation. Therefore, they do not provide a straightforward way of evaluating entire deliberative systems and they do not account for the preconditions of society-wide deliberation.

On the level of systemic deliberation André Bächtiger and John Parkinson (2019) generally distinguish between additive and summative approaches:

> "The first approach is to think that a democratic system gains a deliberative quality when it features institutions that generate strictly defined deliberation at critical points of the system. Moreover, it assumes that the more deliberation there is in the system's component parts, the higher the deliberativeness of the entire system. The second approach is rather different: it is to think that deliberativeness is a quality that emerges from the proper working of the parts of a democratic system, no part of which need be fully deliberative (or fully democratic) on its own"
>
> (Bächtiger and Parkinson 2019: 104)

I believe we need a combination of both approaches. On the one hand, we need to ensure that a range of necessary preconditions for public deliberation are in place such as education

[24]For an overview over the Discourse Quality Index and VisArgue see Marco Steenbergen et al. (2003) and Valentin Gold et al. (2016) respectively.

and (social) security. These are themselves not instances of deliberation, hence they might be overlooked by the additive approach, but sit well with the summative one. On the other hand, and contrary to the summative approach, we also need lively forums and spheres where public deliberation takes place and can be practiced as well as observed. Otherwise, it is improbable that a public of citizens actually develops well-considered preferences on social order.
Now, to my knowledge there is no concrete model of systemic public deliberation for purely epistemic purposes that we could turn to here.[25] Therefore I propose the following framework of the necessary requirements for enabling citizens to have well-considered preferences on social order:

Preconditions of Systemic Public Deliberation

1. **Education:** Recall that, as Shawn Rosenberg (2014) and Owen and G. Smith (2015) have pointed to, individual citizens need to be taught to be deliberators. Effectively, they need an education that provides them with knowledge on the nature and workings of past and present social orders, as well as with certain skills (debate, critically reflection, group discussion). The goal thereby is not to turn all children into public intellectuals, but at least into "standby" public reasoners[26], who have the capability to follow public deliberation and express themselves if they feel that their concerns are excluded.

2. **Security:** Citizens have basic needs for stability, health, certain consumption goods and social inclusion. These should be secured, otherwise citizens are unlikely to have the patience and attention for lengthy debates. This is not only a matter of being fed and comfortable in order to have a discussion, but also of not being afraid. Because, as we know from psychological research, fear and insecurity make people susceptible to the rhetoric of agitators and self-proclaimed strong leaders. (F. Cohen et al. 2014)

Forums of Systemic Public Deliberation

3. **Forums of Public Deliberation:** Besides the listed preconditions for systemic public deliberation, there also need to be actual forums where the forceless force of the better argument can come do bear. But of course, it is difficult to image that, in complex societies, everything is discussed in one single forum. A more realistic picture is that of a "middle democracy" where a whole range of different forums (courts, parliaments, mini-publics,

[25] Jürgen Habermas (2009: 160) provides a model, but he does so for different purposes. Thus he does not provide a purely epistemic model of systemic public deliberation.

[26] The notion of "standby public reasoners" is inspired by Erik Amnå (2014), who speak of the phenomena of standby citizens; "citizens who only appear passive, and in reality are prepared for political action, should circumstances warrant." (Amnå and Ekman 2014: 262)

science, activist- networks) specialize in different kinds of deliberations.(Gutmann and D. Thompson 1996) This division of labor, creates pockets of special knowledge and expertise, as well as different standards of deliberation, fitted to the specific purpose of the forum in question.(Habermas 2006: 415) Nevertheless, the division of labor between different forums serves the overall deliberative quality well, as long as all forums are inclusive and open. More precisely, they need to be inclusive to all perspectives, produce public outputs and be open to public criticism.

4. **A Public Sphere:** The final building block of systemic public deliberation is a uniting, moderated public sphere. A central hub that brings together information and insights from all the more specialized forums before the public of citizens. (Maia; Rousiley 2018) This overarching forum is usually facilitated by mass media journalism. (Habermas 2009: III.8) The technologies of mass media provide the possibility to have such a public sphere in large societies. Journalism on the other hand acts, ideally, as a moderator, breaking down and relating complex pieces of information from the different forums and other sources. This moderation is governed by its own deliberative standards of neutrality, sincerity, respect and openness for criticism. (J. B. Thompson 1995) I take it that the research on mini-public shows that such a moderation is necessary to keep public deliberation from degrading into the mere amplification of the eloquent, the entertaining and the powerful.[27] Moderation is further necessary in order to integrate the different forums and other perspectives (e.g. advocacy groups). Essentially the public sphere integrates and recognizes different forums, groups and perspectives. The division of labor between different forums only works if individual forums are recognized, trusted and interrelated in the public sphere.[28]

Now, this sketch of systemic public reason leaves many details open. Some of them are filled in below in *Table 1*, where I propose how this model could be translated into measurable items. Crucially, all of these proposal are not meant as a model for deliberative democracy, but merely as a collection of systemic requirements for the formation of well-considered preferences on social order.

[27] On the importance of moderated deliberation see (James S. Fishkin 2018: 192) and (Grönlund, Herne, and Setälä 2015).

[28] So, for example, the scientific community relies on being trusted as an authority on matters of fact. This trust is confirmed by the public recognition of scientific findings in the public media and by public demands for other forums, e.g. politicians in parliament, to recognize these findings. Conversely, the scientific community may receive public criticism for not being exclusive or biased, e.g. because professorships are occupied mostly by white males or because some study was funded with cooperate money. The scientific community then has to respond to this criticism in order to maintain trust.

[29] Obviously, this indicator falls way short of establishing what you might call "deliberative education". Ideally there would be an indicator also for the content of education, i.e. for whether children are taught to engage in

Table 5.1: Systemic Public deliberation

Components	**Sub-Components**	**Specifications (Indicator)**
Education – Do individuals enjoy sufficient education and does it reflect deliberative ideals?	Enrollment	Share of population having received public education (OECD: Enrollment rates)
	Higher Education	Share of the population with at least upper secondary education (OECD: Upper secondary education)
	Deliberative Education	Independence from political indoctrination (Freedom House: D3. CIVIL LIBERTIES)[29]
Security – Do individuals enjoy basic rights and securities?	Individual Rights	Do individuals enjoy basic rights granting some personal autonomy? (Freedom House: Personal Autonomy And Individual Rights)
	Social Security	Existence of social protection floors for income, health and pensions (ILO: Decent Work Indicators for Social Security)
Forums of Public Deliberation – What is the deliberative quality of the different forums of public deliberation (e.g. in the polity, the judiciary, academia and civil society) ?	Counsel of ministers	Is this forum recognized in the public sphere? (Counted mention in mass media)
		Is participation restricted based on social or biological traits? (Expert judgment, or V-Dem: Exclusion[30])
		Deliberative quality of the recognized forum (DQI[31]; VisArgue[32])
	Higher court	Same (as Civil Society)
	University	Same (as Civil Society)
	NGO	Same (as Civil Society)
Public Sphere – Are the different forums and citizens related by an overarching, deliberative sphere?	Inclusion	Absence of exclusion (V-Dem: All indicators on exclusion)
	Freedom of the press	Existence of a variety of independent, critical media outlets (Freedom House: Freedom of the press index)
	Freedom of expression	Extension of freedom of expression from the media to the individual (V-Dem: Freedom of expression index)
	Media consumption	Variety of independent sources and degree of professional journalism (Reuters Institute Digital News Report)

5.2.2 Acts of Agreement and Ownership

Now that we have an understanding of the systemic conditions of well-considered choice, we turn to the task of identifying normatively meaningful acts of agreement – a task which essentially consists in integrating the idea of participation as agreement and our normative model. To this end, we firstly need to recall the kind of affirming stances and psychological states carved out in the normative model. That is, we need to recall the kind of acceptance and endorsement of social order that resulted from different kinds of reasons and ownership attachments. Having these things in mind we can then, secondly, identify acts of participation that could plausibly reflect such acceptance and endorsement.

Pragmatic Agreement and the Minimum Point

In *Section 4.2.1*, we came across a pragmatic kind of agreement. Specifically, I claimed that model individuals will accept a social order that, in their eyes, is a mutually beneficial compromise, all things considered. This compromise is at least a case of pragmatic acceptance because individuals are likely to prefer a different order, but nevertheless accept the given order because it allows them to reap the fruits of large-scale cooperation. They also know that any possible social order is likely to be a compromise anyway.

Essentially, such pragmatic acceptance gives us the assurance that we are past the minimum point of preference to the state of nature and thus within the range of justification. That is, individuals who accept social order as a mutually beneficial compromise do not get their most preferred order and may feel ambivalent about the entire set of norms they accept to live by. However, they clearly prefer having this social order over having no order at all.

This kind of pragmatic acceptance is signaled by voluntary and well-informed acts of participation. If, for instance, two people were to sit down and play a game of chess, we do not know much about their motives and preferences. Maybe for both players, chess is their most preferred game. Or, playing chess is really just favored above some alternative, say playing checkers, by player 1, whereas player 2 would prefer checkers but could not convince player 1. Nonetheless, player 2 prefers playing chess over not playing anything at all, thus she agreed to playing chess. It could even be the case that both would prefer checkers over chess, but as it happens, chess is the only game they have available. The point of these scenarios is that the only thing we know when somebody sits down to play a game of chess under normal

productive group discussions about common concerns. Perhaps such an indicator could eventually be provided by expert judgements.

[30]V-Dem is the Varieties of Democracy project (https://www.v-dem.net/).

[31]DQI is the Deliberative Quality Index. (Steenbergen et al. 2003)

[32]VisArgue is a method of measuring the deliberative quality of discussions. (Gold, Hautli-Janisz, and Holzinger 2016)

circumstances is that she prefers playing chess over not playing chess.
Conclusively, agreement as participation in its simplest form allows for the identification of norms that are at least past the minimum point of preference to the state of nature. Nevertheless, in respect to the normative model, we eventually want to know more. We want to know which of the three stages within the entire range of justification has been reached. Thus, we will further need to identify more specific signals of participation.

Let us begin with the first stage of justification, inhabited by what I called "oppportunists" in *Subsection 4.2.3*. As you might recall, opportunists accept their social order for mostly pragmatic reasons, but they do not have an intrinsic motivation to care for the norms they live by. Thus, our opportunists are likely to behave like instrumentally rational agents who only contribute, if they expect a reward for doing so. Therefore, acts of participation by opportunists will need to be moderately incentivized. I say "*moderately* incentivized", because opportunists do not need to be forced to participate. They have their own pragmatic reasons in favor of social order. Still, they are primarily looking out for themselves and try to avoid unnecessary costs. Hence, a moderate level of policing and nudging will generally be sufficient to stabilize a social order between individuals who consider said order to be overall beneficial.

What kind of real-world behavior can signal pragmatic acceptance of social order? I made some suggestions in *Table 5.2* at the end of this subsection. The table lists several potential participatory acts, signaling agreement. Perhaps compliance with the law is the most obvious example. The basic idea is that, since detecting the violation of a social norm usually results in a sanction, this should be sufficient to motivate high levels of compliance amongst opportunists. Thus, in moderately enforced orders, compliance signals at least pragmatic acceptance of said order under the above specified conditions of voluntary and well-informed choice.

Signaling Psychological Ownership

Now let us consider social order with ownership. To this end, recall from *Section 4.2.3* that establishing ownership attachments to a norm means that one has sufficient personal reasons to identify with that norm in question. Therefore it is of inherent value, just as any other positive aspect of the self. The main functional difference implied by this identification with social order is that individuals with ownership endorsement will have intrinsic motivational resources to follow, enforce and improve the norms in question. As a consequence, ownership attachment may motivate *citizenship behavior*, i.e. voluntary, pro-social behavior, irrespective of obvious rewards or payments. Hence, individuals with ownership attachments to norms will tend to be genuine citizens, intrinsically motivation to enact, enforce and improve "their" norms. I have called such citizens with ownership endorsement "stakeholders" to highlight

that they see themselves as non-exclusive owners of the norms they live by.

What kind of real-world behavior can signal ownership endorsement of social order? Here I also list some suggestions in *Table 5.2* at the end of this subsection. Generally, what we are looking for in this context are participatory behaviors that involve moderate costs or effort, rather than benefits to the individual citizen. I say "moderate costs" because we are not talking about any heroic sacrifice here. Rather, stakeholders are expected to invest some effort into the protection, maintenance and improvement of their stake, in this case their social order. Thus, we can expect stakeholders to participate even if this participation comes at a moderate cost, such as informed voting.

Signaling Collective Psychological Ownership

Having social order with collective ownership has two components. First, all individuals share a group identity. Second, all have significant personal reasons motivating ownership attachments to their social order. Since both aspects are common knowledge, all individuals understand each other as members of the same group and claim their social order to be "our" order. I called such citizens "citoyens" in *Subsection 4.2.3*. Essentially, they are stakeholders of social order who have evolved to the level of being collective owners with a collective good perspective.
What kind of real-world behavior can signal ownership endorsement of social order? Again, I list some suggestions in *Table 5.2* below. Generally, the citoyen not only takes responsibility for her stake in social order, but also for social order from a collective perspective. Thus, such citizens are likely to take responsibility for group concerns without expecting a direct reward or repayment. In many respects, the citoyen is likely to be a good *leader* as we understand it today: A person who takes responsibility for the concerns and proceedings of her community and pursues the collective good without seeking material benefit or power. Thus we are looking for leadership in the sense of participation that implies taking responsibility for the collective in the absence of obvious incentives for doing so, such as voluntary engagement in political parties, NGOs or holding an unpaid public office.

Consider *Table 5.2* for a listing of acts of agreement ("sub-components") and respective operationalizations ("specification"). The listing is guided by the goal of identifying types of acts that can be observed on a large scale and fed into an index of justified social order. Therefore, although there are many ways to extend the table with other types of acts, the difficulty is to find things that can be counted and ideally have already been counted and stored in an available database.

Table 5.2: Acts of Agreement

Components	**Sub-Components**	**Specification (Indicators)**
Opportunists – Accept the existing order, avoid extra costs, require incentives to participate	Compliance	Level of compliance in face of moderate enforcement (World Bank: Rule of law)
	Party membership	Party membership where it is voluntary and incentivized
Stakeholders – Endorse the existing order, bear extra costs involved in participation	Voluntary Voting	Voter turnout where voting is not mandatory or incentivized (OECD: Voter turnout)
	Party membership	Party membership where it is voluntary and involves costs rather than benefits
	Legal protest	Acts of protest, covered by the existing political rights
Citoyens - Endorse the existing order, bear extras costs, have group identity, maximize group utility	Civil Society	Participation or membership in non-governmental organizations and institutions (V-Dem: Engagement in political associations; Civil society participation index)
	Party engagement	Party engagement where it is voluntary and involves costs rather than benefits
	Honorary public office	Holding of an unpaid public office

5.2.3 Acts of Disagreement and Defeaters

So far, my thinking about a testing conception of Political Equilibrium has focused on the idea of signaling agreement through acts of participation. Nevertheless, the opposite idea of signaling disagreement through acts of non-participation is also an interesting avenue. The value of the perspective of disagreement consists in an extension of our empirical perspective. This extension is motivated by the problem that those who do not participate are systematically overlooked by our focus on agreement as participation. Consider for instance the possibility that in a given population the levels of agreement we measure are actually countered by significant and unobserved levels of disagreement repsective social order. Surely also the latter should count for something. Hence, counting signals of disagreement also seems highly relevant for the overall degree of justifiedness of a given social order.
In this subsection we consider acts that signal the opposite of acceptance and endorsement of social order: disagreement in the sense of alienation, outright rejection of a norm or social order as a whole. Not surprisingly, in the case of signaling disagreement there are also several different types of acts that might qualify as meaningful signals. Let us consider at least some of the more pertinent ones: alienation, civil disobedience and boycotts as well as political unrest.

Alienation

We begin with the consideration of the exact opposite of agreement as participation: disagreement as non-participation. Just as in the case of agreement as participation the main problem here is that we do not know why a particular person does not participate. So the critical task here consists in identifying a criterion for meaningful disagreement, signaled by political absenteeism. Not surprisingly, there is a long-lasting debate in political science about how to categorize, interpret and evaluate disengagement from politics. (Amnå and Ekman 2014)
For our purposes and in light of the normative model, we may focus on a particular notion of disengagement from social activities that is often presented as the exact opposite of ownership attachment: alienation. (Schacht 2013; McBride 2015; Lafont 2020) Correspondingly, several conceptions of measurements of *political* alienation have been suggested. The following account seems to be the most fitting antonym to ownership:

> "To be politically alienated is to feel a relatively enduring sense of estrangement from existing political institutions, values and leaders. At the far end of the continuum, the politically alienated feel themselves outsiders, trapped in an alien political order; they would welcome fundamental changes in the ongoing regime. By

> contrast, the politically allegiant feel themselves an integral part of the political system; they belong to it psychologically as well as legally."
>
> (Citrin et al. 1975: 3)

Accordingly, measures of political alienation could establish whether and to what extent individuals are excluded from higher stages of justified social order, i.e. justified social order with ownership. This is not to say that alienated individuals are outside of the range of justification. They may still pragmatically *accept* the given order. But in terms of their personal reasons, i.e. in terms of what they want for themselves as the persons they are, they reject this order. Thus they are very far from the ideal scenario of being able to *endorse* the given order as *their own* order. Which is why I consider political alienation an obvious component of a measure of disagreement in coherence with our normative model.
As in the previous subsections, I provide a table (*Table 5.3*) with possible sub-components and indicators at the end of the subsection.

Boycott and Disobedience

Apart from political alienation, there are more specific expressions of disagreement. As you might expect from the Rosa Parks example in *Section 4.2.2*, what I have in mind are civil disobedience and boycotts. Civil disobedience is a public, conscientious and non-violent breach of a norm – whereby non-violence, or rather what it means precisely, is the most contested of these three defining aspects. (Kimberley 2017) However, the non-violence aspect of civil disobedience is of no importance in the context of this section. That is to say that no matter how problematic acts of violence may be from a different perspective, they certainly are no less eligible as candidates for signaling disagreement.
The crucial aspect for our purpose of identifying appropriate signals of disagreement is the fact that individuals are willing to personally bear the costs of norm violation in order to communicate their disagreement. Publicly demonstrating the willingness to bear these costs and potentially being condemned an outsider or outlaw is what makes acts of civil disobedience viable candidates for signaling disagreement. This is because the act in question is a deliberate and costly step outside of the order in question. In contrast to regular political dissent, which is a move within the space of political freedom and participation explicitly granted by (democratic) social order. To distinguish the two cases, let us call the kind of strong disagreement we are interested in here a "rejection" of social order in contrast to disagreement that is merely a dissent from an overall acceptable social order.
Boycotts overlap with civil disobedience in that they also target (directly or indirectly) some specific norm or set of norms which are the object of rejection. Here you may think of a call

to boycott some election as an example. In contrast to civil disobedience, boycotts need not be illegal. Here we should also be careful to distinguish acts that are a political move within a given order (e.g. boycotting some candidate) from a genuine rejection of it (e.g. boycotting an election).
Because civil disobedience and boycotts usually have a fairly specific target, I take that they are a plausible measure for the existence of what I have called "defeaters" in the normative model. These are norms that are outright unacceptable to some and thus prevent these individuals from accepting or endorsing social order as an overall beneficial compromise.

Political Unrest

Besides signaling disagreement by means of non-participation, there are of course also more explicit and straight forward ways of signaling disagreement. Here you might first think of protest. Protest, however, is also ambiguous within the framework of the normative model and the kind of proceduralism discussed in the preceding section. This is because, while protesters often do voice explicit dissent respective to some norm, protest itself is also a regular form of participation within many social orders. Thus, regular protest often does not put the protesting individual outside of the range of justification. More precisely, a protester does not necessarily send a clear signal that the norm in question is a defeater to her, or that she does not prefer the existing order at large to the state of nature. Simply put, protest, especially in democratic orders, is often a political move within the given order and not a signal of rejection of said order.
But of course, there is also the phenomenon of illegal protest. This happens when protesters step outside of the order in place by disobeying standing norms and authorities. This is usually a signal of rejection. Essentially the message is that individuals who engage in illegal protest do not think that their dissatisfaction can not be appropriately addressed within the given order. If many people express this message at the same time, we are confronted with political unrest, i.e. illegal mass mobilization that may lead to violence or even riots.
Political unrest can be like civil disobedience on a large scale. This is the case where large amounts of people converge on considering a particular norm or sets of norms outright unacceptable. But in other cases the people are rising up in order to defeat an entire order that they consider oppressive. They do not want the government or some laws to be changed. They want "the system" to change. Such demands can appear in any political system. However, perhaps the most typical case is an uprising against an autocratic regime which claims the authority of an entire political order.

Consider *Table 5.3* for a listing of acts of disagreement ("sub-components") and respective

Table 5.3: Acts of Disagreement

Components	**Sub-Components**	**Specification (Indicator)**
Alienation – Disagreement as estrangement from social order	Radicalization	Participation in an anti-system opposition movement (V-Dem: CSO anti-system movements)
	Mistrust	Measures of trust in institutions, government and others (OECD: Trust in government and Trustlab-data; European Bank: Life in transition survey)
Political Unrest – Disagreement as illegal mass mobilization against the existing order	Illegal protest and riots	Violent or non-violent but illegal mass mobilization (World Bank: Political Stability and absence of violence/terrorism)
Disobedience and Boycotts – Disagreement as acts of public non-participation or norm violation	Procedural boycotts	Explicit non-participation in an election or other political procedures (The Polity Project: Bycotts in Polity5d)
	Civil disobedience	Violation of a specific norm / sets of norms as a means of protest

operationalizations ("specification"). Recall that the problem here is to find types of acts which, on the one hand, fit the theory and, on the other hand, can and ideally have already been counted.

Now that we have identified some relevant signals of agreement and disagreement in coherence with the normative model, we can turn to the final task of this inquiry: Integrating the identified acts and conditions of systemic public deliberation into an index of justified social order. But before doing so, one clarification is in order: The acts of agreement and disagreement identified above are to be understood as inputs to an aggregate measure of justification, not as a measure of the character of individual people. Effectively, we do not really know what is going on in the individual mind when someone is performing one of the identified acts. Therefore, it would be a mistake to say that someone is a "stakeholder", because she has participated in general elections or in some form of legal protest. Of course, I claim that having personal reasons in favor of one's social order and developing psychological ownership does actually matter in the individual mind. But many other things matter as well. It might for instance be the case that the person in question only participates because she has internalized a norm of participation. We cannot possibly control all factors that might influence individual behavior. Particularly if we are, as in the case of building indices, working with field data instead of

laboratory data. None of these indicators and acts will clarify why some individual does what she does. The claim is rather that having different kinds of reasons and psychological attachments respective to social order does make a measurable difference *on the aggregate level*. This is what an index of justified social order would try to measure.

5.2.4 Toward an Index of Justified Social Order

Why is it a good idea to construct an index of justified social order? There are, of course, *three* reasons in its favor. Firstly, by translating the outcomes of the normative model into an index, we specify what they could mean in terms of observable social states. This translation in form of an operationalization is a well-established way to relate fairly abstract normative theorizing to empirical reality.[33] Doing so simultaneously renders a theory practically relevant in virtue of the resulting measurements and meaningful in virtue of specifying abstract theorizing in terms of observable items. One disadvantage of translating normative theory into an empirical measure is the loss of generality and added controversy about the proper operationalization. This is because the definition components and selection of respective items is rather a matter of considered judgement than of deduction. (Munck and Verkuilen 2002: 7-15)

Secondly, an index of justified social order has the benefit of allowing us to judge instances of social order independent of the existence of democratic procedures. As pointed out in the first chapter, the idea of justified social order and democratic order are closely related in Western political thought. However, an index of justified social order does not require the existence of democratic procedures such as popular voting or voting in parliament. Meaningful acts of agreement or disagreement can take many forms. Thus the applicable range of our testing conception expands beyond that of typical democracy indices and relates to the current interest of political scientists with the issue of deliberation and legitimacy in non-democratic regimes.[34] Further, there is an ongoing debate about potential improvements of existing procedures, such as representation and decision-making by lot, which defies the standard model of modern representative democracy. (G. Smith 2009; Saunders 2010) In light of these research agendas and debates, a normative and comparative measure that is independent of specific ("democratic") procedures is highly desirable.

Thirdly, an index of justified social order allows us to comparatively evaluate where we are in respect to the ideal specified by our normative model. Specifically, the score of the index would

[33] For a fairly recent example of such a translation from normative theory to empirical reality, consider the *Democracy Barometer*. (Bühlmann et al. 2012)

[34] For an example of deliberation in a non-democratic context see Baogang He and Hendrik Wagenaar (2018). For some discussion of legitimacy in non-democracies see Bruce Gilley (2009) as well as Alexander Dukalskis and Johannes Gerschewski (2017) and other entries in the same issue of *Contemporary Politics* on legitimacy in autocracies.

allow us to compare different orders or the same order at different points in time. From this data we could infer how well a given mechanism of norm-selection in some society is actually performing.

A Two-Sided Index of Justified Social Order

As I see it, the discussion in this section does not necessarily lead to the construction of one specific index. This is because, firstly, the discussion has revealed a multidimensional object of measurement, which could be looked at from different sides – e.g. from the side of agreement or from the side of disagreement. Secondly, there are qualitative differences to different acts of agreement (e.g. acceptance vs. endorsement) and different acts of disagreement (e.g. dissent vs. rejection). Thirdly, there is no strictly deductive route from theory to operationalization. Rather, the construction of components, the determination of the aggregating rules and choice of indicators involve many considered judgments.

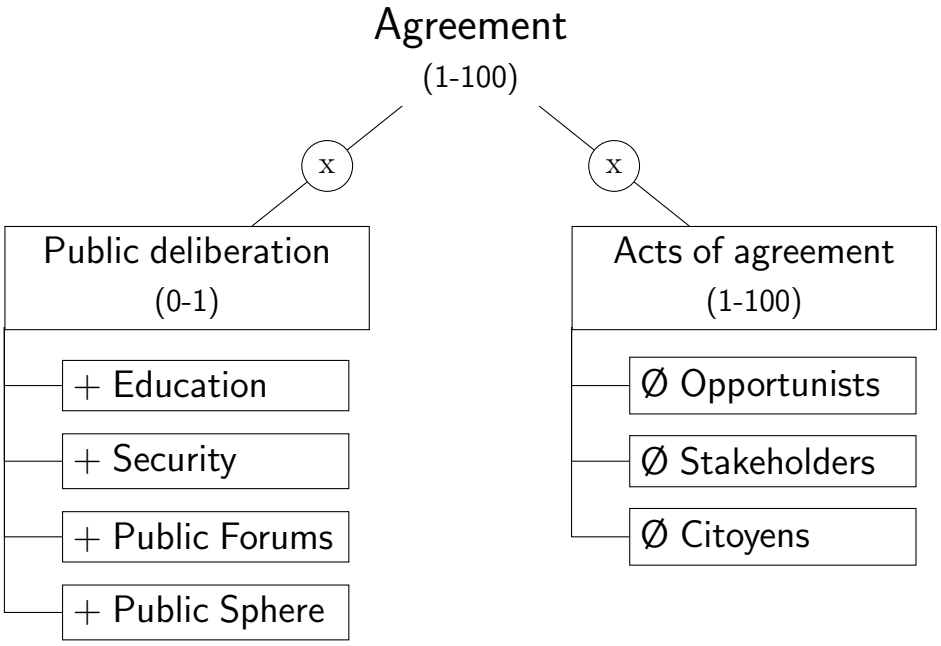
For these reasons I consider the following proposal not as an attempt to present the one and final answer to the question of how to measure justified social order. Rather, it is meant as a first take at a complex task that provides a proof of concept and a construction kit for further, more elaborate attempts.

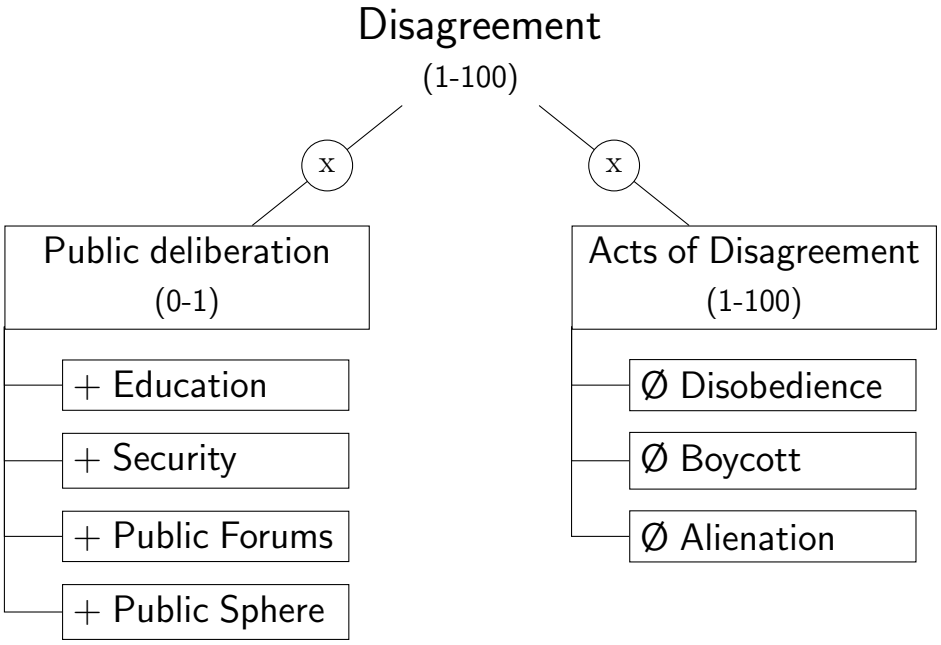
What is the basic idea behind the index? In this section we have discussed three dimensions so far: Systemic public deliberation, acts of agreement and acts of disagreement. The first basic assumption is that acts of agreement and disagreement only matter to the extent that they correlate with the first dimension of systemic public deliberation. The reason for this is simply that if conditions of systemic public deliberation do not obtain, we do not know whether individuals have well-considered preferences and thus we do not know whether their signals of agreement or disagreement are of normative significance. This assumption suggests that the dimension of systemic public discussion should weight the dimensions of agreement and disagreement.

The second assumption is that acts of disagreement should be subtracted from acts of agreement. Despite being distinct observances, both types of acts are of direct relevance for the degree of justification in that acts of agreement add to the degree of justification, while acts of disagreement subtract from it. These assumptions suggest a two-sided design, combining an aggregate measure of agreement and disagreement. In more detail, and following the approach used in the *Polity* index[35], the idea is to calculate the overall score of justified social order by subtracting a weighted measure of disagreement from a weighted measure of agreement. Whereby the weighting is done by multiplying the respective score of agreement and

[35]The *Polity* index is produced by subtracting an aggregate measure of autocracy from an aggregate measure of democracy. (Marshall and Gurr 2020)

disagreement with the score of systemic public deliberation. This design expresses the intuition that acts of agreement and disagreement are equally relevant to the overall degree of justification and that these acts only count to the degree that they are expressed under conditions of systemic public deliberation.
For a better illustration of these arguments, consider a concept tree of the weighted measure of agreement and disagreement.





In both concept trees, the left-hand side represents the dimension of systemic public deliberation. This measure is derived by adding up its four components: education, security, public forums and public sphere. I further suggest to take a note from Marc Bühlmann et al. (2012: 134) and include a penalty for disequilibrium in subcomponent scores. Adding up the scores of the four components with a penalty for disequilibrium reflects two considerations. One, having more of one component is generally better than having less from an overall perspective. Two, having a high value for one component cannot compensate for a low value of a different component. For example, having a very high score in the public sphere component while also having a very low score in education might mean that a lot of individuals are excluded from effectively participating in the public sphere in the first place. These individuals and their concerns remain excluded no matter how high the quality of deliberations in the public sphere. Thus, more balanced scores for both components would be preferable and disequilibrium should result in a reduction of the overall score. In order to achieve this, Marc Bühlmann et al. (2012) suggest using the Arkustangens function as an aggregation rule. Sebastian Jäckle et al. (2012) criticize this method as being too complex and increasing the problem of artificial variance in the measure. Alternatively, they suggest using the geometric mean as an aggregation rule. We do not need to settle this matter here.

The right-hand side of the concept tree consists of aggregate measures for acts of agreement and acts of disagreement respectively. On this side, we are faced with the problem of strong interdependencies between the different components because the different acts measured in each component could be performed by the same individuals. Thus, particularly "noisy" individuals or groups who signal their agreement or disagreement in several ways are likely to be counted several times. Averaging rather than summing up the different components should help mitigate the problem, although it will not eliminate it altogether. This means our measure will probably have a tendency to overestimate acts of agreement and disagreement. But this is not necessarily a problem. To see this, consider that in the case of acts of agreement the overestimation comes in a systematic way that reflects the different levels of justification identified by the normative model. "Opportunists" for example are likely to be counted only one time because they do not engage in acts of participation that require ownership attachments. "Stakeholders" on the other hand are likely to be counted twice because they engage in the same acts of participation as "opportunists" and they engage in acts that require individual ownership attachments. By the same logic, "citoyens" are likely to be counted three times because they are likely to engage in all three kinds of participatory acts. However, the fact that there is only a certain likelihood to be measured more than once and that the different components are averaged out strongly limits the effect of this kind of over-representation. Further, its existence reflects the consideration of the normative model that ownership attach-

ments demarcate a higher levels of justification and thus should indeed count more. So there is a theoretical justification of a systemic over-representation of acts of agreement that signal individual, and more so, collective ownership attachment.

Eventually, the aggregate scores of agreement and disagreement will be weighted by the aggregate score of systemic public deliberation. Finally, the weighted score of disagreement is then subtracted from the weighted score of agreement in order to receive the overall score of justified social order.[36] At this point, overestimations of acts of agreement and disagreement may somewhat cancel each other out.

Unfinished Business

None of the above is meant to suggest that we are close to having an actual index of justified social order. Much more needs to be done. How much more precisely can be specified in reference to the helpful heuristic provided by Gerardo Munck and Jay Verkuilen (2002: 4), who divide up the construction of an index into the three tasks of *conceptualization*, *measurement* and *aggregation*. In this section we have been mostly occupied with fulfilling the task of conceptualization by discussing the three dimensions (systemic public deliberation, acts of agreement and acts of disagreement) and deriving a concept tree. Nevertheless, the conceptualization is arguably not complete until it is reflected in the complete index. Further, I have made some suggestions toward specifying the rules of aggregation, yet a complete mathematical model is still lacking.

Significant work remains to be done in respect to the task of measurement; i.e. the selection of adequate indicators and data sources. What I have provided in Tables 1, 2 and 3 is not much more than a sophisticated brainstorming. As you can see, some sub-components are still lacking indicators and data sources. Also, even in cases where they are provided, it still remains to be determined whether the suggested sources are an adequate basis for a compatible and valid data set. Realistically, I expect that a first functional version of an index of justified social order will be a much simpler version of what I have depicted here. To give an example, consider that in Table 1 I suggest measuring the deliberative quality of different forums in civil society as well as the political, judiciary and academic sphere. First of all, note that I have not specified how many or how often different instances of deliberation in these areas need to be measured. Second of all, even though I do suggest measuring the deliberative quality of specific discussions here, to my knowledge there is no national, let alone international, database for such measurements. Thus, even if automated analyses of transcribed discussion are available,

[36] Obviously, it does not make a numerical difference whether the weighting of the scores is done before or after the disagreement score is subtracted from the agreement score. The reason why I suggest doing it beforehand is that it might be interesting to compare the weighted scores of agreement and disagreement separately between different orders.

collecting the required data would entail significant costs. Therefore I assume that, for the first version of the index, less specific proxies for the quality of these forums will have to suffice.
In respect to the task of aggregation, there is also much work to be done on the level of sub-components. Specifically, plausible minimum and maximum levels and the relative weight of different sub-components need to be determined. For example, one way to get a comparative scale would be to think of counted acts of agreement and disagreement relative to the overall population. Then you might find that, in a given year, 70 percent of society X participated in a popular vote (agreement), whereas 2 percent participated in politically motivated street riots (disagreement). How do we weigh these two different types of acts against each other? Should we simply subtract the 2 percent from the 70 percent? Intuitively, this seems wrong because rioting is a much stronger signal than voting. Nevertheless, considering each individual act of equal importance – thus adhering to a kind of *One person one vote!* principle – might be the most straight-forward approach.
Essentially the crucial goal in determining the precise mathematical model must be that the score of justified order is plausible respective to different levels of agreement and disagreement. That is to say, the overall score should increase with an increase in signaled agreement or a decrease in signaled disagreement and conversely. If this goal is achieved, the overall score can at least serve as a way of tracking the progress within one community. Whether the index is also a plausible comparative measure will need to be tested by plugging in actual data from different societies. It should then become apparent how well the index handles rather obvious cases. So for instance, it would be odd if the index produced dramatically different scores between societies that are fairly similar, according to other measures, or if an order with oppressive leadership, battling its own citizenry in the streets, would receive a relatively high score of justified social order. A more refined benchmark could perhaps be supplied by studies of constitutional approval or deliberative polls, as discussed in *Subsection 3.3.2*.

In spite of these remaining challenges, I hope to have shown that the construction of an actual index of justified social order is a viable possibility. Thus it can be used to complete the third step of Embedded Constructivism by translating PE into social reality. Furthermore, I believe having an index of justified social order would be desirable because it could provide a measure of good social order beyond democracy.

5.3 Concluding Remarks *Chapter 5*

We have arrived at the end of the final chapter of this inquiry. As expected, here I provide an overview over the main arguments and claims put forward in the chapter.

1) In order to complete the third step of Embedded Constructivism, relating the outcomes of the normative model back to social reality, I offer two complementary strategies in this chapter. One, specifying a mechanism for an actual society to establish justified social order understood as Political Equilibrium. Two, a testing conception in the shape of an index of justified social order for testing the level of justification in existing orders.

2) Beginning with the former, the political equilibrium mechanism (PEM) has to fulfill several tasks: Facilitate critically and well-reflected individual preferences, balance group identity and individuality, provide the possibility of self-justification, ensure the selection of eligible norms, and maximize ownership attachments to social order.

3) Since the theory of Political Equilibrium (TPE) is about social norms that are stable for the right reasons, it is obvious that the PEM must include a process of reasoning well in community. How to implement such a procedure is typically discussed under the heading of deliberative democracy. In reflecting upon the relation between TPE and deliberative democracy I focus on the macro perspective of systemic deliberative order and mini-publics.

4) Overall, there are two conclusions resulting from these reflections on the connection between TPE and deliberative democracy. One, a regime of deliberative democracy is the crucial component for any PEM because it is the only plausible way that a broad range of citizens has well-reflected preferences on social order such that justified norms are selected or upheld. Two, there are things relevant to the PEM that go beyond the idea of public deliberation.

5) Therefore, taking notes from Jeremy Waldron (1993) and Gerald Gaus (2016), I argue for a wider proceduralism, which includes, besides a framework of public deliberation, a regime of rights and freedoms, mechanisms of political decision-making as well as space for diversity and experiments in living and in polity.

6) A further and recurrent theme of these preliminary reflections on the PEM is that, from the perspective of theory, we should remain neutral on specific institutional setups. Thus, many things remain open-ended and eventually need to be specified by more local theories, (social) science, politics and citizens of a given society.

7) To complement the PEM and further specify the practical meaning of Political Equilibrium, the second main consideration of the chapter is the testing conception in the form of an index of justified social order. The basic idea is that agreement and disagreement

to a social order can be meaningfully signaled by certain acts of participation or explicit non-participation, given conditions of systemic public deliberation.

8) The first step in specifying the testing conception consists in specifying the core components of systemic public deliberation. Here, I propose a model consisting of two preconditions of public deliberation (education and security) and two kinds of forums where public deliberation actually takes place: One, a set of different forums facilitating division of labour, and two, an integrative, moderated public sphere facilitated by mass media journalism.

9) The second step in specifying the testing conception consists in specifying types of acts that could signal meaningful agreement. In respect to our normative model, I argue that acceptance of an order is signaled by voluntary, moderately incentivized acts of participation (e.g. compliance), whereas ownership endorsement attachment is signaled by voluntary and moderately costly acts of participation (e.g. voting or party engagement).

10) The third step in specifying the testing conception consists in specifying types of acts that could signal meaningful disagreement. Here I argue that disagreement can be signaled by non-participation in the form of political alienation and boycotts, or by explicit violations of standing norms in the form of civil disobedience and illegal protest.

11) The fourth and final step in specifying the testing conception consists in proposing the structure of the actual index of justified social order. Here I suggest that the aggregate scores of agreement and disagreement should be weighted by the aggregate score of systemic public deliberation and that we then calculate the overall score of justified social order by subtracting the disagreement score from the agreement score.

At the very beginning of this chapter I said that its goal is to clarify what it would mean for any actual society to realize the ideal specified in theory, i.e. the ideal of Political Equilibrium as specified in our normative model. In order to achieve this goal, we reflected upon the idea of a political mechanism, selecting norms from the range of justified social order and the idea of a testing conception. We have pursued both avenues without reaching an endpoint. Essentially, we are still lacking a comprehensive understanding of an appropriate political mechanism and functioning index of justified social order.
Nevertheless, I believe even these incomplete reflections allow us to achieve the goal of this chapter. Specifically we gained considerable insights into what it would actually look and feel like to live under a more or less justified social order. One constant in this respect throughout this chapter has been the focus on public deliberation. This is not surprising because the

core ideal of Political Equilibrium is a social order that is stable for the right reasons. Of course one could stumble upon any ideal state by accident, but besides this unlikely scenario, any systematic attempt to get there involves high-quality individual and collective reasoning. Achieving justified order and knowing that this is the case crucially depends on the presence of public deliberation in education, in science, in the judiciary, in civil society, in government and in the overarching public sphere of mass media.

Further, we have learned that achieving justified social order reflects in the way we feel and act. For instance, we may accept the norms we live by despite them being very far from our most preferred norms, simply because we recognize the circumstances of politics and the benefits of stable order. Thus, we abide by these norms most of the time. More ideally, we recognize that the norms we live by reflect many of our own values, which causes us to endorse them as part of ourselves. This more ideal scenario significantly diminishes the tension between the desire to live by self-chosen norms and the reality of a given social order. It should also lead to active citizenship, ranging from simple forms of participation such as voting, all the way to passionate public engagement in the name of some collective good.

Conversely, we have also learned about the face and feel of unjustified social order. This obtains where, in spite of due consideration of the benefits of stable order and the circumstances of politics, social order or some of its components strike people as wholly unacceptable, alien or oppressive. And these sensations of rejection should be visible in acts of withdrawal, disobedience or uprising.

All of this is to say that we have made considerable progress in this chapter with regards to the third step of Embedded Constructivism – the integration of normative theorizing and social reality. Still, there are many questions and tasks that remain unattended. But this is as far as we can proceed within the limits of this inquiry and in many respects also as far as normative social theory should try pressing on alone.

Conclusions

In this final section, I reflect upon the answers given to the original research questions, add some clarifications and point us to open questions for further research. For an overview of all the intermediate steps leading up to this point, I refer the reader to the summaries at the end of each chapter.

Questions and Answers

In *Chapter 1* I posed two guiding questions. Let us consider the respective answers successively. The first question reads as follows:

(1) Under the circumstances of justification, how can social order be justified to each individual governed by its social norms?

Furthermore, I provided the "justification principle of social norms" (JPN) that specified what it would mean to answer this question:

> **JPN**: A social norm N is justified to an individual i in society S governed by that norm to the extent that N being a positive norm in S is coherent with i's preferences, given that
>
> 1) i has formed well-considered preferences on social order.
> 2) N being a positive norm in S is strictly preferred by i to having no social norm governing the domain of N in S.
> 3) i is at liberty to openly reject the JPN in S.

In developing my Theory of Political Equilibrium (TPE), I argued in *Chapter 4* that norms that constitute a mutually beneficial compromise can satisfy the JPN. A compromise is characterized by being favored in light of pragmatic and personal reasons. That is to say, nobody gets their most preferred option, but everybody gets something they want for themselves which renders the entire package desirable overall. This line of thinking implies that there is a whole

range of relations between individual reasons and social norms that fall into the category of justification. Thus there is a "range of justification". On the lower end of the spectrum, we find pure modus vivendi orders that are barely accepted for only pragmatic reasons under the lamentable conditions some group of individuals happens to find itself in. The upper limit of the range of justification is the highly unlikely scenario that everybody can agree on the most preferred set of norms in light of their personal reasons. The more likely case is that social order is favored for a mixture of pragmatic and personal reasons and is thus a genuine compromise where nobody gets their most preferred norms.
Further, I have argued that the ideal way of satisfying the JPN within this framework would be to achieve a compromise with "ownership". A social order as a compromise with ownership denotes the ideal scenario where the set of norms in question is favored by people's personal reasons to the extent that they identify with these norms and recognize them as "*my*" or "*our*" norms. This ideal scenario moves us closer to a sensation of self-legislation and to efficient social order with the potential for advanced forms of cooperation.
In coming to these conclusions in *Section 4.2*, I have simply assumed that conditions (1) and (3) of the JPN are fulfilled. Upon exiting the normative model and getting to more practical matters, we have seen that these conditions as well as the ideal of a compromise with ownership are only achievable under conditions of public deliberation. That is, only under conditions of public deliberation is it plausible that individuals have well-considered preferences on social order, entertain ownership attachments for the right reasons, and can evaluate and reject the JPN and TPE.

In *Subsection 3.2.3* I stated the modified and final version of the second research question as follows:

(2*) How would some given community have to be ordered such that the ideal of justified social order can be systematically pursued by its inhabitants?

The answer provided in *Chapter 5* has two parts. One is the argument for a wide proceduralism, including public deliberation as a crucial component for selecting norms that are stable for the right reasons, and a range of further aspects that are relevant to reasonable political decision-making in a diverse society (i.e. a regime of rights, a decision-making mechanism, and space for diversity in living and in polity). The other part of the answer is that a society could test the justifiedness of its own order, given that it already has institutionalized conditions of systemic public deliberation (i.e. deliberative education, security, specialized spheres of deliberation, and a uniting public sphere).

Lessons and Doubts

I am fairly confident about my criticism of HCM stated in *Chapter 2*. This confidence is not due to a misguided belief in the infallibility of my critical arguments. It rather stems from the observation that, one, John Rawls and Gerald Gaus themselves have downplayed the role of their hypothetical choice models in later writings and that, two, many others have expressed a substantial dissatisfaction with the hypothetical turn. The following statement by James Fishkin is a pointed example:

> "What is not in doubt, at least for me, is that other efforts to express a normatively relevant hypothetical, based on a decision process, have gone far down the road toward abstracting completely from the actual voices of real people under real conditions. The Rawlsian journey began with an early article "Outline of a Decision Procedure for Ethics", which posited only modest impartiality requirements to abstract from the information in actual life. However, progressive refinements eventually yielded a process which shielded the decision maker from virtually all the particulars of actual life. Having spent years in a previous academic life writing within a Rawlsian frame, I eventually concluded that despite its enormous fruitfulness as a theoretical perspective, it yielded a dead end for decision-making, even for fundamental first principles. Even slight differences among assumptions in the original position, all intuitively plausible, lead to starkly contrasting first principles. This conclusion is not original with me, but it helps explain the change of direction of my work from the theory of justice to the theory—and practice—of deliberative democracy."
>
> (James S. Fishkin 2018: 195)

So even if my critical argument proves deficient in some way, I am fairly certain that its overall conclusion expresses a genuine lesson in normative social theory. Namely that the hypothetical turn in social contract theory has lead to an impasse and that models about reasons agents *would have* do not enlighten us about the reasons citizens *do have*. In other words, Dworkin's worry about hypothetical agreement reappears in respect to hypothetical, reason-revealing choice. (Dworkin 1975; D'Agostino, G. Gaus, and Thrasher 2017)

I am also fairly confident that Embedded Constructivism is a plausible approach to normative theorizing that avoids several problems of HCM. More specifically, I think starting out with a descriptive account of the object of theorizing is crucial. Theorists highlighting the *re*constructive aspects of their theory often do this in principle. Embedded Constructivism is about doing so more explicitly and systematically by separating the questions of whether

the empirical model is correct and whether the normative model fits the empirical model. This separation allows for a systematic discussion of the empirical claims involved in the construction and of the fit between the normative and the empirical model. Further, Embedded Constructivism takes notes from how modeling works in social science. The core insight here is that "modeling" is not of much use if its outcomes are not related back to social reality in a systematic way. Thus, the ideals and prescriptions constructed in the normative model need to also be related back to social reality in two ways. First, they need to be translated into guidelines for actual societies in order to be *meaningful*. Second, they need to be addressed to a public of citizens who can asses whether they are *correct*.

My showcase for Embedded Constructivism has been the Theory of Political Equilibrium (TPE). One of the more general lessons from TPE is that we should pay more attention to the psychological dimension of normativity when professing normative social theory. Individual and social normativity exits in our minds in the form of reasons, beliefs, expectations, schemata and scripts, norms, emotional reactive attitudes, psychological attachments and perhaps other things. Normative theorizing ultimately has to be about this normative reality, if it is to be illuminating and guiding for actual social beings.
A more technical lesson from TPE is that, while the social choice perspective of preference aggregation is very helpful in understanding the problem of deep and reasonable pluralism, it is of no assistance in constructing guiding ideals.

I have several doubts in respect to TPE. One concern pertains to the completeness of the empirical model. Perhaps I have overlooked something important and even I have not, future research may still motivate major revisions of the empirical, and consequently also of the normative model. Another concern is that the normative model rests on some speculative presumptions. Namely that reflecting on personal reasons favoring a norm can indeed lead to psychological ownership attachments to that norm. This hypothesis may still be falsified. Or it might turn out that it is true, but that there are several ways how reasoning about norms can lead to a psychological attachment to or internalization of said norms. Hence, as we gain more knowledge, some major revision may have to be incorporated.

> "Add to this the further fact that knock-down, watertight philosophical arguments are always in very short supply; that, in principle, all of one's premises require defence, and that, in practice, not all of them can receive it; and we surely have reason to anticipate other, better work in this area, especially of a more specialized nature."
>
> (Shafer-Landau 2003: 9)

Further Research

TPE is in many ways incomplete. It is explicitly incomplete in terms of a general theory of the good society and social order. Thus, similar to the notion of justice, my account of justification is but one of many virtues of social institutions, and should not be confused with an all-inclusive vision of a good society. (Rawls 1963: 73)

TPE is also obviously incomplete in respect to my initial reflections on the Political Equilibrium Mechanism and the testing conception of justified social order. I do believe that these reflections show that TPE can be systematically related to social reality and that doing so goes a long way toward rendering the normative model practically relevant and meaningful.

Nevertheless, on several occasions we have come up against fascinating questions that cannot be answered within the limits of this inquiry. This includes a theoretical question raised and still debated, in the field of deliberative democracy: What is systemic public deliberation? It also includes the empirical question of whether and under what conditions psychological ownership for norms can be caused by reasoning and more specifically by publicly deliberating about norms.

In respect to Embedded Constructivism I see two obvious subsequent questions. One of them is asking whether Embedded Constructism could be restated as a more general approach to normative theorizing. I have stipulated Embedded Constructivism specifically as an alternative to HCM and in respect to the fairly abstract question of justified social order. Thus, it is still an open matter whether Embedded Constructivism is a plausible approach in other contexts. Intuitively, at least some modifications would have to be made for an application to less abstract objects of theorizing. That is, I do not think that we always need open-ended ideals. If the object of theorizing is some more specific social practice or community, more specific principles may be reconstructed and the results may also apply to social reality in a straightforward manner. This at least seems to be the case in more applied theorizing and ethics.

Another follow-up question in respect to Embedded Constructivism is whether this approach can be explicated as a genuine methodology. In *Chapter III* I have stressed the importance of a *systematic* embedment of normative theorizing. Nevertheless, what counts as "systematic" has remained underdeveloped (with the exception of the construction of an index, but I do not hold that all normative theorizing has to lead to some index). Therefore, Embedded Constructivism so far is at best a methodological approach, but not a genuine methodology – something that allows us to reliably converge on *correct* outcomes. But it sure is desirable to have such a methodology for normative theorizing.

One last question I want to save from the abyss of oblivion is the place for *informal* norms in TPE. In *Chapter 5* I bracketed this issue because thinking about an explicit mechanism

for norm selection just as the idea of an index of justified social order naturally relates to the realm of formal norms. Nevertheless, I believe it could well be worth the effort to think about how agreement as participation could play out in informal scenarios.

This truly concludes my reflections. For now I step aside and await your reply, hopefully given in the arena of public deliberation under an order of public reason.

Bibliography

Amnå, Erik and Joakim Ekman (2014). "Standby Citizens: Understanding Non-Participation in Contemporary Democracies". In: *European Political Science Review* 6.2, pp. 261–281.

Arrow, Kenneth J. (1950). "A Difficulty in the Concept of Social Welfare". In: *The Journal of Political Economy* 58.4, pp. 328–346.

– (1951). *Social Choice and Individual Values*. New York.

Austin, John (1832). *The Province of Jurisprudence Determines*. London.

Avey, James B. et al. (2009). "Psychological ownership: Theoretical extensions, measurement and relation to work outcomes". In: *Journal of Organizational Behavior* 30, pp. 173–191.

Axelrod, Robert (1984). *The Evolution of Cooperation*. New York.

Bächtiger, André and John Parkinson (2019). *Mapping and Measuring Deliberation: Towards a New Deliberative Quality*. Oxford.

Baer, Markus and Graham Brown (2012). "Blind in one eye: How psychological ownership of ideas affects the types of suggestions people adopt". In: *Organizational Behavior and Human Decision Processes* 118.1, pp. 60–71.

Baier, Kurt (1995). *The Rational and the Moral Order: The Social Roots of Reason and Morality*. Chicago.

Banks, James A. (2008). "Diversity, Group Identity, and Citizenship Education in a Global Age". In: *Educational Researcher* 37.3, pp. 129–139.

Baron, Robert Steven (2005). "So Right It's Wrong: Groupthink and the Ubiquitous Nature of Polarized Group Decision Making". In: *Advances in Experimental Social Psychology* 37, pp. 219–253.

Basu, Kaushik (2015). "The Republic of Beliefs: A new Approach to Law and Economics". In: *Policy Research Working Paper* 7259, pp. 1–54.

Batalha, Luisa Maria et al. (2019). "Psychological Mechanisms of Deliberative Transformation: The Role of Group Identity". In: *Journal of Public Deliberation* 15.1.

Bell, Daniel (2020). *Communitarianism*. URL: https://plato.stanford.edu/archives/fall2020/entries/communitarianism/.

Benhabib, Seyla (1990). "Communicative Ethics and Current Controversies in Practical Philosophy". In: *The Communicative Ethics Controversy*. Ed. by Seyla Benhabib and Fred Dallmayr. Cambridge MA / London, pp. 330–369.

Bicchieri, Cristina (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge, p. 260.

– (2017). *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. New York.

Bicchieri, Cristina and Erte Xiao (2009). "Do the Right Thing: But Only If Others Do SoNo Title". In: *Journal of Behavioral Decision Making* 22.2, pp. 191–208.

Binmore, Kenneth (2005). *Natural Justice*. Oxford / New York.

Black, Duncan (1948). "On the Rationale of Group Decision-making". In: *Journal of Political Economy* 56.1, pp. 23–34.

Bolino, Mark C. (1999). "Citizenship and impression management: Good soldiers or good actors?" In: *Academy of Management Review* 24.1, pp. 82–98.

Bowles, Samuel and Herbert Gintis (2004). "The evolution of strong reciprocity:cooperation in heterogeneous populations". In: *Theoretical Population Biology* 65, pp. 17–28.

– (2011). *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton.

Brennan, Geoffrey et al. (2013). *Explaining Norms*. New York.

Brown, Graham and Sandra L. Robinson (2007). "The dysfunction of territoriality in organizations". In: *New horizons in management. Research companion to the dysfunctional workplace: Management challenges and symptoms*. Ed. by Janice Langan-Fox, Cary L. Cooper, and Richard J. Klimoski, pp. 252–267.

Brown, James Robert and Yiftach Fehige (2019). *Thought Experiments*. URL: https://plato.stanford.edu/archives/win2019/entries/thought-experiment/.

Buchanan, James M. and Gordon Tullock (1962). *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. Ann Arbor.

Bühlmann, Marc et al. (2012). "Demokratiebarometer: ein neues Instrument zur Messung von Demokratiequalität". In: *Zeitschrift fur Vergleichende Politikwissenschaft* 6.1, pp. 115–159.

Carbonara, Emanuela (2017). "Law and Social Norms". In: *The Oxford Handbook of Law and Economics: Volume 1: Methodology and Concepts*. Ed. by Francesco Parisi. New York / Oxford, pp. 466–482.

Carr, Craig L. (1990). "Tacit Consent". In: *Public Affairs Quarterly* 4.4, pp. 335–345.

Carruthers, Peter (2006). *The Architecture of the Mind*. Oxford / New York.

Cheney, Dorothy L. and Robert M. Seyfarth (1990). *How monkeys see the world: Inside the mind of another species*. Chicago.

Christiano, Thomas (2008). *The Constitution of Equality: Democratic Authority and its Limits*. Oxford / New York.

Chwe, Michael Suk-Young (2001). *Rational Ritual: Culture, Coordination,and Common Knowledge*. Princeton / Oxford.

Cialdini, Robert B., Raymond R. Reno, and Carl A. Kallgren (1990). "A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places." In: *Journal of Personality and Social Psychology* 58.6, pp. 1015–1026.

Citrin, Jack et al. (1975). "Personal and Political Sources of Political Alienation". In: *British Journal of Political Science* 5.1, pp. 1–31.

Cohen, Florette et al. (2014). "Fatal Attraction: The Effects of Mortality Salience on Evaluations of Charismatic, Task-Oriented, and Relationship-Oriented Leaders". In: *Psychological Science* 15.12, pp. 846–851.

Cohen, Joshua (1989). "Deliberation and Democratic Legitimacy". In: *The Good Polity*. Ed. by Alan Hamlin and Philip Pettit. Oxford.

Cooter, Robert (1998). "Expressive Law And Economics". In: *The Journal of Legal Studies* 27.S2, pp. 585–607.

Cunningham, George B. (2005). "The importance of a common in-group identity in ethnically diverse groups". In: *Group Dynamics: Theory, Research, and Practice* 9.4, pp. 251–260.

D'Agostino, Fred (2013). "The Orders of Public Reason". In: *Analytic Philosophy* 54.1, pp. 129–155.

– (2018). "How Can We do Political Philosophy?" In: *Cosmos + Taxis* 5.2, pp. 29–37.

D'Agostino, Fred, Gerald Gaus, and John Thrasher (2017). *Contemporary Approaches to the Social Contract*. URL: https://plato.stanford.edu/archives/sum2017/entries/contractarianism-contemporary/.

Darwall, Stephen, Allan Gibbard, and Peter Railton (1992). "Toward Fin de siècle Ethics: Some Trends". In: *The Philosophical Review* 101.1, pp. 115–189.

Davis, Taylor and Daniel Kelly (2018). "Norms, not moral norms: The boundaries of morality do not matter". In: *Behavioral and Brain Sciences* 41.e101, pp. 18–19.

Diekmann, Andreas et al. (2014). "Reputation Formation and the Evolution of Cooperation in Anonymous Online Markets". In: *American Sociological Review* 79.1, pp. 65–85.

Downs, Anthony (1957). "An Economic Theory of Political Action in a Democracy". In: *Journal of Political Economy* 65.2, pp. 135–150.

– (1962). "The Public Interest: Its Meaning in a Democracy". In: *Social Research* 29.1, pp. 1–36.

Dryzek, John S. and Hayley Stevenson (2011). "Global democracy and earth system governance". In: *Ecological Economics* 70.11, pp. 1865–1874.

Dukalskis, Alexander and Johannes Gerschewski (2017). "What autocracies say (and what citizens hear): proposing four mechanisms of autocratic legitimation". In: *Contemporary Politics* 23.3, pp. 251–268.

Dworkin, Ronald (1975). "The Original Position". In: *Reading Rawls: Critical Studies on Rawls' A Theory of Justice*. Ed. by Norman Daniels. Oxford, pp. 16–53.

Ekins, Paul (1992). "Towards a Progressive Market". In: *Banking People*. Ed. by Udo Reifner and Janet Ford. New York, pp. 43–54.

Elster, Jon (1992). *Local Justice: How Institutions Allocate Scarce Goods and Necessary Burdens*. Cambridge.

Enoch, David (2015). "Against Public Reason". In: *Oxford Studies in Political Philosophy, Vol. 1*. New York, pp. 112–142.

Estlund, David M. (2008). *Democratic Authority, A Philosophical Framework*. Princeton / Oxford.

Faillo, Marco, Daniela Grieco, and Luca Zarri (2013). "Legitimate punishment, feedback, and the enforcement of cooperation". In: *Games and Economic Behavior* 77.1, pp. 271–283.

Fehr, Ernst and Urs Fischbacher (2004). "Third-party punishment and social norms". In: *Evolution and Human Behavior* 25.2, pp. 63–78.

Fischer, Stefan (2018). *The Origin of Oughtness: A Case for Metaethical Conativism*. Berlin.

Fishkin, James S and Robert C Luskin (2005). "Experimenting with a Democratic Ideal: Deliberative Polling and Public Opinion". In: *Acta Politica* 40.1284–298.

Fishkin, James S. (1991). *Democracy and Deliberation: New Directions for Democratic Reform*. New Heaven / London.

– (1995). *The Voice of the People: Public Opinion and Democracy*. New Heaven / London.

– (2018). "Response to Critics: Toward the Reform of Actually Existing Democracies". In: *The Good Society* 27.1, pp. 190–210.

Fishkin, James S. et al. (2010). "Deliberative Democracy in an Unlikely Place: Deliberative Polling in China". In: *British Journal of Political Science* 40.2, pp. 435–448.

Freeman, Samuel (2007). *Justice and the Social Contract: Essays on Rawlsian Political Philosophy*. Oxford / New York.

Friedman, Milton (1953). "The Methodology of Positive Economics". In: *Essays in Positive Economics*. Chicago, pp. 3–43.

Friedman, Milton and Rose Friedman (1980). *Free to Choose*. London.

Galston, William A. (1982). "Moral Personality and Liberal Theory: John Rawls's "Dewey Lectures"". In: *Annual Review of Political Science* 10.4, pp. 492–519.

Gaus, Daniel (2013). "Rational Reconstruction as a Method of Political Theory between Social Critique and Empirical Political Science". In: *Constellations* 20.4, pp. 553–570.

Gaus, Gerald (2011). *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. Cambridge.

– (2016). *The Tyranny of the Ideal: Justice in a Diverse Society*. Princeton / Oxford.

– (2017). "Is Public Reason a Normalization Project? Deep Diversity and the Open Society". In: *Social Philosophy Today* 33.1, pp. 27–52.

– (2018). "The Complexity of a Diverse Moral Order". In: *The Georgetown Journal of Law and Public Policy* 16.S, pp. 645–679.

Gert, Bernard and Joshua Gert (2020). *The Definition of Morality*. URL: https://plato.stanford.edu/archives/fall2020/entries/morality-definition/.

Gettier, Edmund L. (1963). "Is Justified True Belief Knowledge?" In: *Analysis* 23.6, pp. 121–123.

Gilley, Bruce (2009). *The right to rule: How states win and lose legitimacy*. New York.

Gintis, Herbert (2008). "Punishment and Cooperation". In: *Science* 319.5868, pp. 1345–1346.

Gold, Valentin, Annette Hautli-Janisz, and Katharina Holzinger (2016). "VisArgue: Analyse von politischen Verhandlungen". In: *Zeitschrift für Konfliktmanagement* 19.3. Ed. by Uta Hüttig, pp. 98–99.

Griep, Yannick, Timothy Wingate, and Carmien Brys (2017). "Integrating Psychological Contracts and Psychological Ownership: The Role of Employee Ideologies, Organisational Culture and Organisational Citizenship Behaviour". In: *Theoretical Orientations and Practical Applications of Psychological Ownership*. Ed. by Chantal Olckers, Llewellyn van Zyl, and Leoni van der Vaart. Cham, pp. 79–102.

Grofman, Bernard (2004). "Downs And Two-Party Convergence". In: *Annual Review of Political Science* 7, pp. 25–46.

Grönlund, Kimmo, Kaisa Herne, and Maija Setälä (2015). "Does Enclave Deliberation Polarize Opinions?" In: *Political Behavior* 37.4, pp. 995–1020.

Gutmann, Amy and Dennis Thompson (1996). *Democracy and disagreement*. Cambridge MA.

Habermas, Jürgen (1984). *The Theory of Communicative Action, Vol. 1*. Boston.

– (1995). "Reconciliation Through the Public use of Reason: Remarks on John Rawls's Political Liberalism". In: *The Journal of Philosophy* 92.3, pp. 109–131.

– (2006). "Political Communication in Media Society: Does Democracy Still Enjoy an Epistemic Dimension? The Impact of Normative Theory on Empirical Research". In: *Communication Theory* 16, pp. 411–426.

– (2008). *Between Naturalism and Religion*. Cambridge / Malden.

– (2009). *Europe: The Faltering Project*. Cambridge / Malden.

Haidt, Jonathan (2001). "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment". In: *Psychological Review* 108.4, pp. 814–834.

Halbig, Christoph (2007). *Praktische Gründe und die Realität der Moral*. Frankfurt (a.M.)

Hamilton, William Donald (1963). "The Evolution of Altruistic Behavior". In: *The American Naturalist* 97.896, pp. 354–356.

Harsanyi, John C. (1953). "Cardinal Utilityin Welfare Economics and in the Theory of Risk-Taking". In: *Journal of Political Economy* 61, pp. 434–435.

– (1955). "Cardinal Welfare, Individualistic Ethics and Interpersonal Comparisons of Utility". In: *Journal of Political Economy* 63, pp. 309–321.

– (1978). "Bayesian Decision Theory and Utilitarian Ethics". In: *The American Economic Review* 68.2, pp. 223–228.

Hayek, Friedrich A. (1935). *Collectivist Economic Planing*. London.

– (1981). "The Flow of Goods and Services". In: *The Collected Works of F. A. Hayek (2012)*. Ed. by H. Klausinger. vol. 8. Chicago.

He, Baogang and Hendrik Wagenaar (2018). "Authoritarian deliberation revisited". In: *Japanese Journal of Political Science* 19.4, pp. 622–629.

Hinsch, Wilfried (2018). "Expectation-Based Legitimacy". In: *Human Rights, Democracy, and Legitimacy in a World of Disorder*. Ed. by Silja Voeneky and Gerald L. Neuman. Cambridge, pp. 97–110.

Hobbes, Thomas (1651). *Leviathan*. London.

Hoffman, Elizabeth et al. (1997). "Preferences, Property Rights, and Anonymity in Bargaining Games". In: *Games and Economic Behavior1* 7.3, pp. 346–380.

Hornsey, Matthew J. and Jolanda Jetten (2004). "The Individual Within the Group: Balancing the Need to Belong With the Need to Be Different". In: *Personality and Social Psychology Review* 8.3, pp. 248–264.

Hotelling, Harold (1929). "Stability in Compitition". In: *The Economic Journal* 39.153, pp. 41–57.

Hume, David (1748). "Of the Original Contract". In: *Essays, Moral, Political, and Literary*. Ed. by Eugene F. Miller (1985). Vol. 50. 2. Indianapolis, pp. 465–487.

Ihara, Yasuo and Marcus W. Feldman (2004). "Cultural niche construction and the evolution of small family size". In: *Theoretical Population Biology* 65.1, pp. 105–111.

Jäckle, Sebastian, Uwe Wagschal, and Rafael Bauschke (2012). "Das Demokratiebarometer: „basically theory driven"?" In: *Zeitschrift fur Vergleichende Politikwissenschaft* 6.2, pp. 99–125.

Kahneman, Daniel (2011). *Thinking Fast and Slow*. New York.

Kant, Immanuel (1781). *Kritik der Reinen Vernunft*. Berlin: Preussische Akademie der Wissenschaften (1900ff.)

– (1785). *Grundlegung zur Metaphysik der Sitten*. Berlin: Preussische Akademie der Wissenschaften (1900ff.)

– (1797). *Die Metaphysik der Sitten*. Berlin: Preussische Akademie der Wissenschaften (1900ff.)

Kelly, Daniel (n.d.). "Two Ways to Adopt a Norm: The (Moral?) Psychology of Internalization and Avowal". In: *The Oxford Handbook of Moral Psychology* ().

Kimberley, Brownlee (2017). *Civil Disobedience*. URL: https://plato.stanford.edu/archives/fall2017/entries/civil-disobedience/.

Klosko, George (1997). "Political Constructivism in Rawl's Political Liberalism". In: *The American Political Science Review* 91.3, pp. 635–646.

Lafont, Cristina (2020). *Democracy without Shortcuts: A Participatory Conception of Deliberative Democracy*. Oxford.

Lamm, Ehud (2014). "Forever united: the co-evolution of language and normativity". In: *The Social Origins of Language*. Ed. by Daniel Dor, Chris Knight, and Jerome Lewis, pp. 267–283.

Landemore, Hélène (2013). *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many*. Princeton.

Lenman, James and Yonatan Shemmer (2012). *Constructivism in Practical Philosophy*. Oxford.

Lessnoff, Michael Harry (1986). *Social Contract*. London.

Lewis, David (1969). *Convention: A Philosophical Study*. Oxford.

Lisciandra, Chiara, Caterina Marchionni, and Alessandra Basso (2017). "Hypothetical models in social science: their features and uses". In: *Springer Handbook of Model-Based Science*. Ed. by Lorenzo Magnani and Tommaso Bertolotti. Dordrecht, pp. 413–433.

Locke, John (1690). *Second Treatise of Government*. London.

Luskin, Robert C. et al. (2014). "Deliberating across Deep Divides". In: *Political Studies* 62.1, pp. 116–135.

Macedo, Stephen (2010). *Why Public Reason? Citizens' Reasons and the Constitution of the Public Sphere*. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1664085.

MacKenzie, Michael K (2016). "Institutional design and sources of short-termism". In: *Institutions for future generations*. Ed. by Iñigo González-Ricoy and Axel Gosseries. Oxford, pp. 24–48.

Mackie, Gerry (2018). "Effective Rule of Law Requires Construction of a Social Norm of Legal Obedience". In: *Cultural Agents Reloaded: The Legacy of Antanas Mockus*. Ed. by Carlo Tognato. Cambridge MA.

Mackie, John Leslie (1977). *Ethics: Inventing Right and Wrong*. London.

Maia; Rousiley (2018). "Deliberative Media". In: *The Oxford Handbook of Deliberative Democracy*. Ed. by Andre Bächtiger et al. Oxford.

Mansbridge, Jane et al. (2012). "A systemic approach to deliberative democracy". In: *Deliberative systems: deliberative democracy at the large scale*. Ed. by John Parkinson and Jane Mansbridge. Cambridge, pp. 1–26.

March, Andrew and Alicia Steinmetz (2018). "Religious Reasons in Public Deliberation". In: *The Oxford Handbook of Deliberative Democracy*. Ed. by Andre Bächtiger et al. Oxford.

Marshall, Monty G. and Ted Robert Gurr (2020). "POLITY5: Dataset Users' Manual". URL: http://www.systemicpeace.org/inscr/p5manualv2018.pdf.

Martinaityte, Ieva, Kerrie L. Unsworth, and Claudia A. Sacramento (2020). "Is the project 'mine' or 'ours'? A multilevel investigation of the effects of individual and collective psychological ownership". In: *Journal of Occupational and Organizational Psychology* 93.2, pp. 302–327.

McBride, Cillian (2015). "Democratic Ownership and Deliberative Participation". In: *Political and Civic Engagement: Multidisciplinary Perspectives*. Ed. by Martyn Barrett and Bruna Zani. London, pp. 109–123.

McGeer, Victoria and Philip Pettit (2002). "The self-regulating mind". In: *Language & Communication* 22, pp. 281–299.

Mercier, Hugo and Hélène Landemore (2012). "Reasoning Is for Arguing: Understanding the Successes and Failures of Deliberation." In: *Political Psychology* 33.2, pp. 243–58.

Mercier, Hugo and Dan Sperber (2017). *The Enigma of Reason*. Cambridge MA.

Miller, David (2003). *Principles of social justice*. Cambridge MA.

Moehler, Michael (2015). "The Rawls–Harsanyi Dispute: A Moral Point of View". In: *Pacific Philosophical Quarterly* 99.1, pp. 82–99.

– (2018). *Minimal Morality: A Multilevel Social Contract Theory*. Oxford.

Müller, Julian (2019). *Political Pluralism, Disagreement and Justice - The Case for Polycentric Democracy*. London / New York.

Munck, Gerardo L. and Jay Verkuilen (2002). "Conceptualizing And Measuring Democracy: Evaluating Alternative Indices". In: *Comparative Political Studies* 35.1, pp. 5–34.

Nagel, Thomas (1974). "What Is It Like to Be a Bat?" In: *The Philosophical Review* 83.4, pp. 435–450.

Neblo, Michael (2010). *Change for the Better? Linking the Mechanisms of Deliberative Opinion Change to Normative Theory*.

Niesen, Peter (2016). "Die politische Theorie des politischen Liberalismus: John Rawls". In: *Politische Theorien der Gegenwart III*. Ed. by André Brodocz and Gary S. Schaal. Opladen Berlin Toronto, pp. 25–63.

– (2017). "Constituent power in global constitutionalism". In: *Handbook on Global Constitutionalism*. Ed. by Anthony F. Lang and Antje Wiener. Oxford, pp. 222–233.

Nijs, Tom et al. (2020). "'This country is OURS': The exclusionary potential of collective psychological ownership". In: *British Journal of Social Psychology* 60.1, pp. 171–195.

Nozick, Robert (1974). *Anarchy, State, and Utopia*. Oxford / Cambridge MA.

O'Driscoll, Michael P., Jon L. Pierce, and Ann-Marie Coghlan (2006). "The Psychology of Ownership: Work Environment Structure, Organizational Commitment, And Citizenship Behaviors". In: *Group & Organization Management* 31.3, pp. 388–416.

Ostrom, Elinor (2000). "Collective Action and the Evolution of Social Norms". In: *Journal of Economic Perspectives* 14.3, pp. 137–158.

Owen, David and Graham Smith (2015). "Survey Article: Deliberation, Democracy, and the Systemic Turn". In: *The Journal of Political Philosophy* 23.2, pp. 213–234.

Ozler, Hayrettin, Abdullah Yilmaz, and Derya Ozler (2008). "Psychological ownership: An empirical study on its antecedents and impacts upon organizational behaviors". In: *Problems and Perspectives in Management* 6.3, pp. 38–47.

Parkinson, John (2006). *Deliberating in the Real World: Problems of Legitimacy in Deliberative Democracy*. Oxford / New York.

– (2018). "Deliberative Systems". In: *The Oxford Handbook of Deliberative Democracy*. Ed. by Andre Bächtiger et al. Oxford.

Patberg, Markus (2014). "Supranational constitutional politics and the method of rational reconstruction". In: *Philosophy and Social Criticism* 40.6, pp. 501–521.

Peng, He (2013). "Why and when do people hide knowledge?" In: *Journal of Knowledge Management* 17.3, pp. 398–415.

Pettit, Philip (1997). *Republicanism: A Theory of Freedom and Government*. Oxford.

Pierce, Jon L. and Liro Jussila (2010). "Collective psychological ownership within the work and organizational context: Construct introduction and elaboration". In: *Journal of Organizational Behavior* 31.1, pp. 810–834.

– (2011). *Psychological Ownership and the Organizational Context: Theory, Research Evidence, and Application*. Cheltenham / Northampton.

Pierce, Jon L., Tatiana Kostova, and Kurt T. Dirks (2003). "The state of psychological ownership: Integrating and extending a century of research". In: *Review of General Psychology* 7, pp. 84–107.

Posner, Richard A (1997). "Social Norms and the Law: An Economic Approach". In: *American Economic Review* 87.2, pp. 365–369.

Postmes, Tom, Gamze Baray, et al. (2006). "The Dynamics of Personal and Social Identity Formation". In: *Individuality and the Group Advances in Social Identity*. Ed. by Tom Postmes and Jolanda Jetten. London, pp. 215–236.

Postmes, Tom and Russell Spears (2005). "Individuality and Social Influence in Groups: Inductive and Deductive Routes to Group Identity." In: *Journal of Personality and Social Psychology* 89.5, pp. 747–763.

Quong, Jonathan (2018). *Public Reason*. URL: https://plato.stanford.edu/archives/spr2018/entries/public-reason/.

R. Kelly, Daniel (2017). "Moral Cheesecake, Evolved Psychology, and the Debunking Impulse". In: *The Routledge Handbook of Evolution and Philosophy*. Ed. by Richard Joyce. New York, pp. 342–358.

Rawls, John (1963). "Constitutional Liberty and the Concept of Justice". In: *Nomos VI: Justice*. Ed. by C. J. Friedrich and John Chapman. New York.

– (1967). "Distributive Justice". In: *Philosophy, Politics and Society, Third Series*. Ed. by Peter Laslett and W.G. Runciman. Oxford.

– (1971). *A Theory of Justice*. Cambridge MA.

– (1980). "Kantian constructivism in moral theory". In: *Journal of Philosophy* 77.9, pp. 515–572.

– (1993). *Political Liberalism*. New York.

– (1997). "The Idea of Public Reason Revisited". In: *The University of Chicago Law Review* 64.3, pp. 765–807.

– (2001). *Justice as Frainress: A Restatement*. Cambridge MA.

Raz, Joseph (1999). *Engaging Reason*. Oxford / New York.

Rosenberg, Shawn W. (2014). "Citizen competence and the psychology of deliberation". In: *Deliberative Democracy: Issues and Cases*. Ed. by Stephen Elstub and Peter McLaverty. Edinburgh, pp. 98–117.

Rousseau, Jean-Jacques (1762). *On the Social Contract; or, Principles of Political Rights*. Amsterdam.

Samuelson, Paul and William D. Nordhaus (2009). *Economics*. Boston.

Saunders, Ben (2010). "POLITY5, Political Regime Characteristics and Transitions, 1800-2018 Dataset Users' Manual". In: *Ethics* 121.1, pp. 148–177.

Schacht, Richard (2013). "Alienation". In: *The International Encyclopedia of Ethics*. Ed. by Hugh LaFollette. Hoboken (NJ), pp. 198–206.

Schelling, Thomas (1960). *The Strategy of Conflict*. Cambridge MA / London.

Schramme, Thomas (2006). *Gerchtigkeit und soziale Praxis*. Frankfurt / New York.

Schwartz, Shalom H., Gian Vittorio Caprara, and Michele Vecchione (2010). "Basic Personal Values, Core Political Values, and Voting: A Longitudinal Analysis". In: *Political Psychology* 31.3, pp. 421–452.

Sen, Amartya K. (1977). "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory". In: *Philosophy & Public Affairs* 6.4, pp. 317–344.

Shafer-Landau, Russ (2003). *Moral Realism: A Defence*. Oxford.

Shank, Daniel B. et al. (2019). "Norm Talk and Human Cooperation: Can We Talk Ourselves Into Cooperation?" In: *Journal of Personality and Social Psychology* 117.1, pp. 99–123.

Skyrms, Brian (1996). *Evolution of the Social Contract*. Cambridge.

– (2016). "Evolution, Norms, and the Social Contract". In: *Arizona State Law Journal* 48.4, pp. 1087–1099.

Smith, Graham (2009). *Democratic Innovations: Designing institutions for citizen participation*. Cambridge.

– (2018). "Mini-Publics and Deliberative Democracy". In: *The Oxford Handbook of Deliberative Democracy*. Ed. by Andre Bächtiger et al. Oxford.

Smith, J. Maynard and George R. Price (1973). "The logic of animal conflict". In: *Nature* 246, pp. 15–18. ISSN: 00280836. DOI: 10.1038/246015a0.

Sommers, Tamler (2016). *A Very Bad Wizard: Morality Behind the Curtain*. 2nd. New York.

Spiekerman, Kai (2015). "Book Review: Explaining Norms, Geoffrey Brennan, Lina Eriksson, Robert E. Goodin and Nicholas Southwood". In: *Economics and Philosophy* 31.1, pp. 175–181.

Sripada, Chandra Sekhar and Stephen Stich (2006). "A Framework for the Psychology of Norms". In: *The Innate Mind, Volume 2: Culture and Cognition*. Ed. by Peter Carruthers, Stephen Laurence, and Stephen Stich. Oxford, pp. 280–316.

Stanford, P. Kyle (2018a). "Moral externalization and normativity: The errors of our ways". In: *Behavioral and Brain Sciences* 41.e119, pp. 34–49.

– (2018b). "The difference between ice cream and Nazis: Moral externalization and the evolution of human cooperation". In: *Behavioral and Brain Sciences* 41.e95, pp. 1–13.

Steenbergen, Marco R. et al. (2003). "Measuring Political Deliberation: A Discourse Quality Index". In: *Comparative European Politics* 1.1, pp. 21–48.

Stemmer, Peter (2008). *Normativität*. Berlin.

– (2016). *Der Vorrang des Wollens, Eine Studie zur Anthropologie*. Frankfurt (a.M.)

– (2017). "Moral, moralisches Müssen und Sanktionen". In: *Deutsche Zeitschrift für Philosophie* 65.4, pp. 621–656.

Stephanopoulos, Nicholas O. and Mila Versteeg (2016). "The Contours of Constitutional Approval". In: *Washington University Law Review* 94.1, pp. 113–190.

Sugden, Robert (1986). *The Economics of Rights, Co-operation andWelfare*. Oxford.

Summers, Jesse S. (2017). "Rationalizing our Way into Moral Progress". In: *Ethical Theory and Moral Practice* 20, pp. 93–104.

Sunstein, Cass R (1996). "Social Norms and Social Roles". In: *Columbia Law Review* 96.4, pp. 903–968.

Swaab, Roderick et al. (2007). "Shared Cognition as a Product of, and Precursor to, Shared Identity in Negotiations". In: *Personality and Social Psychology Bulletin* 33.2, pp. 187–199.

Talisse, Robert B. (2014). "Moral authority and the deliberative model". In: *Philosophical Studies* 170.3, pp. 555–561.

Thompson, John B. (1995). *The Media and Modernity: A Social Theory of the Media*. Cambridge.

Thomson, Judith Jarvis (1985). "The Trolley Problem". In: *Yale Law Journal* 94.6, pp. 1395–1415.

Uhlmann, Eric Luis et al. (2009). "The motivated use of moral principles". In: *Judgment and Decision Making* 4.6, pp. 476–491.

Valdesolo, Piercarlo and David DeSteno (2006). "Manipulations of Emotional Context Shape Moral Judgment". In: *Psychological Science* 17.6, pp. 476–477.

Vallier, Kevin (2011). "Convergence and Consensus in Public Reason". In: *Public Affairs Quarterly* 25.4, pp. 261–279.

– (2018). *Public Justification*. URL: https://plato.stanford.edu/archives/spr2018/entries/justification-public/.

Van Dyne, Linn and Jon L. Pierce (2004). "Psychological ownership and feelings of possession: Three field studies predicting employee attitudes and organisational citizenship behavior". In: *Journal of Organizational Behavior* 25.1, pp. 439–459.

Van Schoelandt, Chad (2018). "Moral Accountability and Social Norms". In: *Social Philosophy and Policy* 35.2, pp. 217–236.

Vanderschraaf, Peter (1995). "Conventions as Correlated Equilibria". In: *Erkenntnis* 42, pp. 65–87.

Vanderschraaf, Peter and Brian Skyrms (1993). "Deliberational Correlated Equilibria". In: *Philosophical Topics* 21.1, pp. 191–227.

– (2003). "Learning to Take Turns". In: *Erkenntnis* 59.3, pp. 311–347.

Vaughn, Karen I. (2013). "Hayek, Equilibrium, and The Role of Institutions in Economic Order". In: *Critical Review* 25.3-4, pp. 473–496.

Verkuyten, Maykel and Borja Martinovic (2017). "Collective Psychological Ownership and Intergroup Relations". In: *Perspectives on Psychological Science* 12.6, pp. 1021–1039.

Waldron, Jeremy (1987). "Theoretical Foundations of Liberalism". In: *The Philosophical Quarterly* 37.147, pp. 127–150.

– (1993). "A Right-Based Critique of Constitutional Rights". In: *Oxford Journal of Legal Studies* 13.1, pp. 18–51.

– (1999). *Law and Disagreement*. Oxford.
– (2007). "Public Reason and "Justification" in the Courtroom". In: *Journal of Law, Philosophy and Culture* 1.1, pp. 107–137.

Wall, Steven (2017). "Is Public Justification Self-Defeating?" In: *American Philosophical Quarterly* 39.4, pp. 385–394.

Weber, Max (1964). *Wirtschaft und Gesellschaft. Grundrisse der Verstehenden Soziologie*. Köln.

Wendt, Fabian (2016). *Compromise, Peace and Public Justification: Political Morality Beyond Justice*. London.

Wenner, Fabian (2013). "Die Idee des demokratischen Prozeduralismus bei Jeremy Waldron". URL: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjDz_jDsqbuAhXllYsKHWZAAp4QFjAAegQIBRAC&url=https%3A%2F%2Fwww.uni-muenster.de%2Fimperia%2Fmd%2Fcontent%2Fkfg-normenbegruendung%2Fintern%2Fpublikationen%2F57_wenner_-_demokratischer_p.

Wyman, Emily (2014). "Language and collective fiction: from children's pretence to social institutions". In: *The Social Origins of Language*. Ed. by Daniel Dor, Chris Knight, and Jerome Lewis, pp. 171–183.