Pluralism, Consensus, and Justification

A Simulation Study on
Overlapping Consensus in Liberal Democracies

Zur Erlangung des akademischen Grades eines
DOKTORS DER PHILOSOPHIE (Dr. phil.)

von der KIT-Fakultät für Geistes- und Sozialwissenschaften des
Karlsruher Instituts für Technologie (KIT)
angenommene

DISSERTATION

von
Richard Lohse

KIT-Dekan: Prof. Dr. Michael Mäs

1. Gutachter: Prof. Dr. Gregor Betz
2. Gutachter: Prof. Dr. Dr. Claus Beisbart

Tag der mündlichen Prüfung: 27. November 2024

Acknowledgments

I would like to express my deepest gratitude to the following individuals for their invaluable support and guidance throughout the development of this thesis.

First and foremost, I am profoundly grateful to my supervisor, Prof. Dr. Gregor Betz, for his continuous encouragement and insightful feedback. Some of the most important ideas in this thesis emerged from our discussions. I especially appreciate that he took time for counsel whenever needed, but never put too much pressure on me. I could not have asked for a better supervisor.

I also wish to extend my heartfelt thanks to the other members of our research team, Prof. Dr. Dr. Claus Beisbart, Prof. Dr. Georg Brun, Sebastian Cacean and Dr. Andreas Freivogel, as well as Sebastian Flick, Dr. Alexander Koch and Noah Werder. Our regular meetings have been nothing but productive and fun. My thesis profited immensely from your input. Not least, the present research would not have been possible without the fantastic Python library developed by Andreas and Basti.

Special thanks are due to my colleagues, Dr. Eike Düvel, Kurt Halter and Dr. Michael Schmidt, as well as the participants of the KIT philosophy research seminar, who generously read early drafts of this thesis. Their thoughtful critiques and supportive remarks were invaluable in refining the final manuscript.

To all of you, I extend my deepest appreciation. This thesis would not have been possible without your contributions. The same applies in a more indirect though not less relevant way to all the teachers, mentors, and friends who supported and profoundly influenced me over the years, including but

not limited to, in alphabetical order: Dr. Christian Barth, Dr. Jochen Briesen, Dr. Stefan Fischer, Dr. Andrea Lailach-Hennrich, Prof. Dr. Thomas Müller, Prof. Dr. Jacob Rosenthal, Dr. Antje Rumberg, Prof. em. Dr. Wolfgang Spohn, Prof. em. Dr. Peter Stemmer, Dr. Verena Wagner, and Prof. Dr. Alexandra Zinke.

Last but not least, I am grateful for the financial support provided, first, by the German Research Foundation (grant 412679086) and, second, by Prof. Dr. Christian Seidel-Saul. Their funding enabled me to conduct the research and complete this thesis.

To Antje Rumberg and Jacob Rosenthal

Deutsche Zusammenfassung (Summary in German)

Viele Gesellschaften sind *pluralistisch*. Das heißt insbesondere, dass sie eine Vielfalt an Weltanschauungen aufweisen. Trotzdem wäre es gut, wenn es unter den Bürgerinnen einen *Konsens* zumindest in grundlegenden Verfassungsfragen gibt. Zudem ist es wünschenswert, dass jede einzelne Bürgerin die Verfassung nicht aus Zwang, sondern aus guten Gründen akzeptiert, d.h. in ihren Meinungen *gerechtfertigt* ist. Diese drei Bedingungen (Pluralismus, Konsens, Rechtfertigung) können in einem Spannungsverhältnis stehen. Insbesondere kann es sein, dass die Vielfalt an Weltanschauungen derart ist, dass ein gerechtfertigter Konsens in Verfassungsfragen unmöglich wird.

Wenn dies der Fall ist, gibt es vier Möglichkeiten, damit umzugehen: Wir können entweder auf Konsens verzichten, auf Rechtfertigung verzichten, Pluralismus ganz abschaffen oder aber Bedingungen schaffen, die zu einem Pluralismus führen, der Konsens und Rechtfertigung nicht im Wege steht. Die ersten drei Optionen, so argumentiere ich, sind mit erheblichen Problemen verschiedener Art verbunden.

Daher ist es von großer Bedeutung, die vierte Option zu untersuchen und Umstände zu finden, unter denen es gleichzeitig Pluralismus, Konsens und Rechtfertigung geben kann. Um es mit dem Begriff von John Rawls zu sagen, müssen wir Umstände finden, unter denen ein *überlappender Konsens* möglich ist. In dieser Arbeit versuche ich, einen Beitrag zu diesem Ziel zu leisten. Dabei konzentriere ich mich auf den erkenntnistheoretischen Aspekt der Aufgabe: Was muss der Fall sein, damit das Rechtfertigungskriterium erlaubt, dass eine Konstellation von Glaubenssystemen sowohl Pluralismus als auch Konsens aufweist? Diese Frage hat bis dato wenig direkte Aufmerksamkeit erfahren.

Meine Methodik zur Untersuchung dieser Frage stützt sich auf formale und komputationale Erkenntnistheorie. Insbesondere kann die vorliegende Dissertation in zwei Teile geteilt werden.

Im ersten Teil, dem *philosophischen* Teil, wenn man so will, stelle ich zunächst die Rawls'sche Idee eines überlappenden Konsens etwas detaillierter dar. Ich betone jedoch, dass meine Arbeit eine strukturelle, der Logik nicht unähnliche Perspektive einnimmt und damit von vielen heiß diskutierten Fragen der Rawls'schen Theorie unabhängig ist. Das heißt insbesondere, dass die Arbeit für Philosophinnen verschiedener Lager von Interesse ist. Anschließend entwickle ich eine Handvoll Definitionen für verschiedene Stadien eines überlappenden Konsenses. Diese Definitionen sind meines Wissens so noch nicht formuliert worden und bilden die Grundlage für die vorliegende und eventuell anschließende Arbeiten. Besonders hervorzuheben ist die Unterscheidung zwischen einem globalen und einem lokalen überlappenden Konsens. Letzterer ist der schwächere Begriff und erfordert lediglich, dass es in einer *Teilgesellschaft* Pluralismus, Konsens und Rechtfertigung gibt. Man kann versuchen, daraus einen globalen überlappenden Konsens zu machen, indem man untersucht, welche günstigen Bedingungen in der Teilgesellschaft herrschen und versucht, diese günstigen Bedingungen auch im Rest der Gesellschaft herzustellen.

Desweiteren entwickle ich in diesem ersten, philosophischen Teil der Dissertation eine Definition des hier relevanten Rechtfertigungsbegriffs. Dieser fußt auf der Methode des Überlegungsgleichgewichts. Grob gesagt gilt ein Meinungssystem als gerechtfertigt genau dann, wenn es das Ergebnis der Anwendung dieser Methode hätte sein können. Bei Anwendung der Methode müssen auch alternative Sichtweisen und Argumente, nicht nur die eigenen, berücksichtigt werden. Die Gesamtheit aller zu berücksichtigenden Sichtweisen und Argumente nenne ich *dialektische Situation*. Ich argumentiere, dass die dialektische Situation jeder Bürgerin mindestens all jene Sichtweisen und Argumente enthält, die in breiter Weise *öffentlich debattiert* werden, z.B. in den klassischen und sozialen Medien, in Parlamenten, etc. Die dialektischen Situationen der Bürgerinnen werden erwartbarerweise einen signifikanten Einfluss auf die Möglichkeit eines überlappenden Konsenses haben. Die vorliegende Dissertation soll diesen Einfluss untersuchen.

Genauer geht es um den Einfluss der inferentiellen Beziehungen zwischen den öffentlich debattierten Weltanschauungen und der Verfassung. Ich stelle die in der Arbeit zu prüfende Hypothese auf, dass nur Weltbilder, die die Verfassung *stützen*, einen überlappenden Konsens bezüglich selbiger möglich machen. Sollte sich diese Hypothese als zutreffend erweisen, könnte dies einen hohen Standard an die öffentliche Debatte stellen. Insbesondere könnte das im schlimmsten Fall bedeuten, dass mit der Verfassung *inkompatible* oder auch nur *neutrale* Weltbilder von der öffentlichen Debatte ausgeschlossen werden müssen. Insbesondere bezüglich letzterer ist das nicht wünschenswert, da äußerst illiberal. Es ist also wichtig, diese Hypothese zu überprüfen.

Im zweiten Teil der Dissertation, dem formalen und komputationalen Teil, schlage ich mathematische Explikationen der verschiedenen Arten von überlappendem Konsens vor. Anschließend stelle ich Design und Ergebnisse einer Simulationsstudie vor, die dazu dient, die o.g. Hypothese zu überprüfen.

Die mathematische Explikation des zuvor definierten Rechtfertigungsbegriffs basiert auf dem formalen Modell des Überlegungsgleichgewichts, das jüngst von Claus Beisbart, Gregor Betz und Georg Brun vorgestellt wurde. Ich stelle dieses vor, passe es an die vorliegende Problematik an und gebe schließlich ein mathematisches Kriterium für Rechtfertigung an. Dieses kann auch von Computern in sog. Simulationen berechnet werden. Anschließend stelle ich verschiedene mathematische Maße für Konsens und Pluralismus vor. Diese sind so entworfen, dass sie der Untersuchung der Frage nach überlappendem Konsens dienen. Die mathematischen Explikationen von Rechtfertigung, Konsens und Pluralism werden nun zu präzisen und in Studien anwendbaren Explikationen der verschiedenen Arten von überlappendem Konsens zusammengefügt.

Anschließend stelle ich eine breit angelegte Simulationsstudie vor, die die Hypothese testen soll. Diese Studie beruht nicht auf empirischen Daten. Stattdessen werden zufällig generierte, künstliche, und relativ kleine Gesellschaften simuliert. Die 'Bürgerinnen' befinden sich je nach Gesellschaft in verschiedensten dialektischen Situationen. Insbesondere liegen verschiedenste Kombinationen von inferentiellen Beziehungen zwischen Weltbildern und Verfassung vor. Für jede Gesellschaft kann mit den zuvor entwickelten

Explikationen untersucht werden, ob ein überlappender Konsens in globalem oder lokalem Sinne vorliegt. Es wird dann ausgewertet, für welche Kombinationen von inferentiellen Beziehungen besonders oft oder besonders selten ein überlappender Konsens vorliegt. Anhand dieser Auswertung kann die Hypothese überprüft werden. Natürlich sind wir letztlich nicht an künstlichen Gesellschaften interessiert. Ich erkläre, warum sich aus den Ergebnissen auch entsprechende Rückschlüsse auf echte Gesellschaften ziehen lassen.

Das Ergebnis der Simulationsstudie ist, grob gesagt, dass die Hypothese für *globale* überlappende Konsense bestätigt oder zumindest nicht falsifiziert wird. Das heißt, dass in der öffentlichen Debatte nur stützende Weltbilder einen globalen überlappenden Konsens möglich machen. Allerdings gilt für *lokale* überlappende Konsense ein niedrigerer Standard. Für solche ist lediglich erforderlich, dass es keine inkompatible Weltbilder in der öffentlichen Debatte gibt. Neutrale Weltbilder hingegen sind nicht hinderlich.

Ich diskutiere auch, welche Konsequenzen sich aus diesen Ergebnissen ziehen lassen. Zunächst ist das Ergebnis für globale überlappende Konsense besorgniserregend, weil es wie erwähnt bedeuten könnte, dass inkompatible und neutrale Weltbilder von der öffentlichen Debatte ausgeschlossen werden müssen. Ich lote aus, inwieweit sich diese illiberale Konsequenz eventuell vermeiden lässt. Das Ergebnis für lokale überlappende Konsense hingegen macht Hoffnung. Wie erwähnt kann man untersuchen, welche Bedingungen in der jeweiligen Teilgesellschaft vorliegen und untersuchen, ob solche günstigen Bedingungen auch im Rest der Gesellschaft herstellbar sind. Ich spekuliere, was diese Bedingungen sein könnten und schlage Folgestudien vor, die diese Spekulation überprüfen. Sollte wir entsprechende günstige Bedingungen finden, lässt sich das besorgniserregende Ergebnis für globale überlappende Konsense möglicherweise vermeiden.

Die vorliegende Dissertation legt den Grundstein für weitere Untersuchungen der hochrelevanten Frage, wie ein gerechtfertigter Konsens in Verfassungsfragen trotz weltanschaulichem Pluralismus möglich ist. Insbesondere legt sie eine Reihe fruchtbarer begrifflicher Präzisierungen vor, die für solche Untersuchungen, insbesondere empirischer Natur, unerlässlich sind. Desweiteren liefert sie mit der durchgeführten Simulationsstudie einen Aufschlag, der weiterführende Forschung anregt.

Contents

Chapter 1

Introduction

Consider the following initially plausible statements:

Pluralism	Societies are often pluralist, i.e. the citizens hold a diversity of worldviews.
Consensus	It would be good if citizens in a society agreed on constitutional essentials concerning the procedure and limits of political decision making.
Justification	It would be good if citizens in a society were justified in holding their beliefs.

There is a tension between these statements: Some pluralist societies exhibit a combination of worldviews that make it unlikely, if not impossible for citizens to agree on constitutional essentials while at the same time being justified in holding their beliefs. To see this possibility, consider the following highly stylised, fictional case of a pluralist society. Suppose around half of the citizens are religious zealots. Of central importance to them, especially their moral beliefs, is a holy scripture which they think is written by God. The scripture gives strict rules regarding virtually every aspect of life. In particular, it is intolerant. It demands that one pursue any trespassing of these rules by others, even if they are not believers themselves. Also, it demands an extreme form of proselytism, i.e. the zealots will go very far to convert others to their faith. Whoever resists conversion is killed. The other half of the citizens are steadfast atheists. They believe in a world devoid of deities and ruled exclusively by the laws of science. Personal

freedom is important to them, they will mostly do whatever they feel like doing and often break the strict rules of the zealots. However, the atheists do agree (amongst themselves) on few basic moral rules, e.g. they agree on "Do not kill another person unless for self-defense" on the grounds of a mutual interest of not being killed. Philosophically speaking, they adopt a broadly contractarian perspective on morality while the zealots embrace a theistic one.

Needless to say, this won't go well. Let's take their combination of worldviews as given and fixed. Then how on earth can they find a common basis for political decision making? The differences between their worldviews run too deep. But perhaps not all hope is lost and a group of zealots or a group of atheists might try to force a consensus despite the deeply opposed worldviews by using the means of propaganda and indoctrination, perhaps paired with the resources of an oppressive authoritarian regime. This might work, but it is clear that, as a consequence, citizens would not be justified in holding their beliefs. That is, their beliefs would not be based on good reasons, reliable processes or whatnot, but on propaganda. As a consequence, it seems that *given* this particular combination of worldviews *either* there will be no consensus on basic political questions *or* citizens do not hold justified belief systems because the consensus is forced by means of propaganda. We cannot have both consensus and justification.

Of course, this is an extreme case and luckily it isn't real. However, there are real cases that bear similarities to this fictional society (see below for examples). For now, note that this outcome is not inevitable for any pluralist society. In particular, there seem to be kinds of pluralism, i.e. combinations of worldviews, such that there *can* be both consensus and justification. Soon I'll discuss German society as one such example. Thus, there are combinations of worldviews such that consensus and justification can go together and there are combinations such that this is difficult if not impossible.

The idea that citizens with differing worldviews might nonetheless justifiedly agree on constitutional essentials was prominently spelled out by political philosopher John Rawls. He called such a constellation of belief systems an *overlapping consensus*: the different worldviews justifiedly overlap on a shared conception of constitutional essentials (Rawls, 1987, 2005; see also Taylor, 1999; Finlayson, 2019). The idea of an overlapping consensus is

the main subject of this thesis. If an overlapping consensus isn't possible, because the pluralism is such that it stands in the way of justified consensus, then there are 4 options:

1. Accept that there is no consensus.
2. Accept that there is no justification.
3. Abolish pluralism altogether.
4. Bring about conditions such that an overlapping consensus is possible, i.e. such that the pluralism doesn't stand in the way of a justified consensus.

Below I argue that the first three options are undesireable. As a consequence, we should investigate what option 4 could like like and whether it is more desireable than the others. The goal of this thesis is to contribute to this task by investigating the crucial question: Under which conditions is an overlapping consensus possible? The remainder of this introduction is structured as follows. First, I lay out why this question is important. That is, I argue that the first three options are undesireable. Second, I explain which aspect of the very broad question will be in focus, namely, the epistemological aspect. I sketch how I use tools from formal and computational epistemology to address it. Third, I give a detailed overview of the structure and content of the present thesis.

First things first, why are the first three options undesireable? Let's start with option 1. Two examples will highlight that a consensus on constitutional essentials is important.

In November 2020 Donald Trump lost the United States presidential elections against Joe Biden. The election and its aftermath was overshadowed by numerous allegations of widespread voter fraud. These allegations were made by government officials, including Trump himself, Republican politicians and ordinary citizens. However, all purported evidence of such fraud, if it was given at all, has been debunked as either harmless or fabricated (see, e.g., Politifact, 2020). That is, Joe Biden won fair and square. Nonetheless, Trump to this day refuses to concede victory to Biden. He has mounted numerous attempts to overturn the election result, including pressuring Georgia election official Brad Raffensberger and firing Chris Krebs,

the director of the federal Cybersecurity and Infrastructure Security Agency (e.g. CNN, 2021; NY Times, 2020). These attempts culminated in a pro-Trump rally near the White House on January 6, 2021, the very day on which the US Congress was to finally confirm Biden's victory over Trump. At the rally, Donald Trump himself gave a speech, urging his supporters to march towards the Capitol to protest against this final confirmation by the Congress (Washington Post, 2021).

And they did. Roughly 10,000 Trump supporters marched onto Capitol grounds. The protests turned into violent riots. Around 1,000 people attacked and broke into the Capitol building in an attempt to stop the confirmation of Biden's victory (NPR, 2022). Ultimately, the rioters were not successful. Even though the process was interrupted, security services were able to regain control and the confirmation resumed later that night (ibid). Joe Biden became 46th President of the United States of America. This violent attack on the stable functioning of democratic institutions was a direct consequence of Trump's and his supporter's refusal to accept the outcome of the elections. Of course, even though this attack is shocking, 1,000 rioters is a small number when compared against over 300.000.000 million US citizens who did not participate in these protests. Still, around *one third* of US citizens believe that the election was stolen. This statistic has been established by different organisations at different points in time (Ipsos, 2021a,b; Monmouth Poll Reports, 2021; Rasmussen Reports, 2022).

Two years later in Iran, on September 14, 2022, Mahsa Amini was arrested for not properly wearing her hijab and for wearing tight pants (Al Jazeera, 2022). The Islamic Republic of Iran has strict rules concerning the dress code for women, e.g. they are required to wear hijabs that completely cover their hair as is required by (their interpretation of) Sharia law. The dress code is rigorously enforced by a vice squad called the Guidance Patrol, a part of the Iranian police force. They're quite busy: In 2014, it gave guidance to 3.6 million Iranians, in addition to taking over 200,000 women to police stations for improperly wearing their hijabs (Parsa, 2016). Mahsa Amini died in a hospital two days after being arrested. Officially, the cause of death was sudden heart failure. However, other women in detention witnessed that she was severely beaten and died of police brutality (Guardian, 2022). Journalist Nilofaar Hamedi publicised the case with her photo of the grieving parents.

This sparked nation-wide outrage. Massive protests spread from Amini's hometown throughout Iran. Supreme Leader Ali Khamenei discredited the protests as caused by foreign nations and tried to brutally crush the protests, resulting in hundreds of deaths, thousands of arrests and at least eight executed death sentences (Iran Human Rights, 2023). During 2023 the protests dwindled, the ruling elite remains entrenched in power (Reuters, 2023). To be sure, police brutality and the compulsary hijab were not the only reasons for the protests. As in the protests in the years before, a general discontent with the theocratic political system and the economic situation formed the backdrop (AP Press, 2022).

I believe that both cases, the attack on the Capitol and the Mahsa Amini protests, show that stability is threatened in the respective society. By 'stability' I here mean the most basic form: absence of widespread social unrest, or at least absence of civil war. In Iran, this is very obvious and pressing. Perhaps stability is not only threatened, but already crumbling away. It might be only a matter of time before the next widespread protests erupt. In the United States, this is less obvious. The attack on the Capitol was not successful. Joe Biden became President and is since able to govern more or less as other presidents before him. But suppose there is a rematch between Trump and Biden in the next presidential elections (both have already announced their candidacy) and Biden wins again. Will the Trump supporters accept that result? Or will there be riots again, perhaps more widespread, like the ones in Iran? Given that one third of US citizens still believe that Biden is not the legitimate president, we simply don't know. Thus, even though US society is currently stable, that stability is threatened.

I hypothesise that the deeper problem in both societies is that there is no consensus on the two most basic questions of living together: First, how should political decisions be made? Second, what are the limits of any political decision making? In many societies, the answer to the first question is 'democracy' (EIU, 2023) and the answer to the second is 'human rights' (HRW, 2023). However, citizens need not only agree on these broad terms, but also on the specifics. Regarding (representative) democracy, they need to agree on all of the relevant details of the election procedure. Regarding human rights, citizens need to agree on the content of these rights and which rights are more important than others in case of conflict. Alternatively, if

some citizens have not thought much about these details, they nonetheless need to generally agree that these things are regulated and handled in a way that is acceptable to them.

In the US, citizens do *not* agree on the specifics of the democratic procedure. In particular, they do not agree on what grounds are sufficient for the 2020 presidential election outcome to be rejected as the result of voter fraud. For one third of the population, some dubious videos on social media and the word of the election loser Trump himself seems to be enough. At least, it is enough for them when it's in their own interest to reject the outcome. Others disagree. This disagreement threatens societal stability, as the attack on the Capitol shows. In Iran, citizens do not agree about whether the government (democratically elected or not) may for religious reasons impose a certain dress code on women. There are those that think that the government may and perhaps should do that. And, as the Mahsa Amini protests show, there are those that think that the government must not do that. Again, this disagreement threatens social stability. Of course, these are just two examples of societal stability being threatened by a missing consensus on basic political questions. It's not hard to find more, just think of the 2023 protests against the judicial reform in Isreal, or the 2023 Brazilian Congress attack.

My point is that a consensus on constitutional essentials is good, because it promotes societal stability. Or put differently, choosing option 2 by sacrificing consensus leads to problems, because it may lead to societal instability. I know of no empirical research confirming this positive correlation between consensus and stability, but the two examples I discussed (and abundant further ones) illustrate that it seems very plausible, almost obvious, that there is such a correlation. Of course, there are differences between the two examples I gave. The Islamic Republic of Iran, being a *de facto* theocracy, can more easily try to maintain stability by using brute force, thus far successfully. The USA, on the other hand, is a democracy and as such has less resources to maintain stability by brute force. In essence, even though a consensus on constitutional essentials is strictly speaking neither necessary nor sufficient for societal stability, it nonetheless *promotes* such stability. In particular, the societal stability in liberal democracies seems to depend rather heavily on a consensus on constitutional essentials, especially when

compared to authoritarian regimes. I conclude that option 1 (accepting that there is no consensus) is undesireable.

Let's skip option 2 for the moment and have a look at option 3, i.e. abolishing pluralism. A society is pluralist in the sense intended here, if it exhibits a high diversity of religions, ideologies, moral doctrines and values, or, to use a catch-all term, a high diversity of worldviews. Of course, societies can be diverse in many other respects as well, I just listed those that most obviously have the potential to impede a justified consensus on basic political questions. Should we fight pluralism in order to maintain consensus and stability? I think there are reasons against doing so. For one thing, many consider diversity of worldviews a *ceteris paribus* good thing. For example, one might simply enjoy living in a diverse and colourful environment, with new experiences and stimulating exchanges. Or one might think that diversity increases productivity, because involving different perspectives enhances our capacities to solve problems (CdV, 2016). For another, more important thing, even if one does not think that pluralism is a *ceteris paribus* good thing, in many societies worldview pluralism is just a brute fact that is not easily changed. Let me explain.

First, the violent way to abolish pluralism would be to separate the people who hold, say, the dominant worldview from the people who don't and then either kill or deport the latter. This is obviously morally wrong, at least for the vast majority of worldviews that actually exist. (Of course, the zealots described above do not have these moral scruples and there might be actual cases that are similar in this respect, think of the terror group Islamic State.) Nonetheless, in some cases the separation of groups with different worldviews might be possible, sensible and morally permissible. For example, if the groups are already spatially separated, then political secession might be the best solution. But these cases are the exception, not the rule. Political secession is not easily achieved and sometimes quite violent, think of Yugoslavia's bloody breakup in the nineties (Silber and Little, 1996).

Second, even if a society is not yet pluralist, it can be morally problematic to try to keep it that way. For one thing, many consider it morally obligatory to admit foreign people into society, e.g. because they are persecuted in theirs, or because of war or natural disaster (for an overview of the arguments, see

Parekh, 2020, ch. 3). For another, to some extent pluralism seems to be the natural outcome of a liberal state protecting freedom of speech, conscience, press, science, and so forth. This is a point John Rawls stresses (2005, p. xxiv). Thus, trying to prevent pluralism might require abolishing these freedoms to some extent which counts as a very serious moral cost in many worldviews. This, again, shows that the problem is more pressing for liberal democracies when compared to more oppressive forms of regime.

Finally, there is the option of trying to persuade citizens, i.e. not using oppressive means like propaganda to change their beliefs, but giving good reasons, *convincing to all*, such that only one worldview remains in that society. I think that this is hopeless. The best indicator for this is the fact that moral philosophers have still not found these decisive reasons, convincing to all, for the one true or most plausible moral theory, even though they have been at it for a while now (for a recent documentation of this disagreement, see Bourget and Chalmers, 2023, cf. the question on normative ethics in table 1). If these bright minds, who dedicate their lives to thinking about these matters, cannot find such reasons, then perhaps we should suppose that there are none. I am not advocating that moral philosophers stop the search, but until they are successful we should, for practical matters, assume that it cannot be done.

At this point you might object that, even though we don't know of decisive reasons for the one true moral theory, most if not all moral theories agree regarding a restricted range of cases, for example, they agree that killing people for mere pleasure is wrong. (Famously, Küng (1990) tries to describe this agreement, for normative contractarian projects in similar vein see Stemmer (2000); Moehler (2018).) Thus, there may be decisive reasons for this kind of 'minimal morality'. Call me skeptical, but let's suppose that there are such reasons. Then these reasons will convince all citizens of this minimal morality. But they will leave open what to think about cases outside this restricted range. Thus, these reasons will not strongly diminish the diversity of worldviews in a society.

Taking these points together, I submit that option 3, i.e. abolishing pluralism, isn't a viable one. At least, this holds for liberal democracies which can't use oppressive means to keep citizens in line with a single worldview.

Let's turn to option 2, i.e. accepting that citizens are not justified in hold-

ing their views. It is obvious that this is epistemically bad. But it is also bad for the well-being of the citizens. For one thing, having justified beliefs is important to us. When someone is charged with having unjustified beliefs, then that person will likely either rebut the charge by insisting that their beliefs are justified, or change their beliefs such that they think them justified. What's unlikely to happen is that the person will answer 'So what?', because having justified beliefs is just not important to them. Thus, since having what's important to us matters for our well-being, option 3 will be detrimental to our well-being. For another thing, suppose that we can bring it about (by whatever means) that citizens hold a diversity of worldviews, but only one conception of constitutional essentials and that the price for this is that many citizens hold *inconsistent* beliefs. For example, in the model case of zealots and atheists living together, we managed to make them agree on a theocratic regime that is to the zealots' liking. The atheists agree on this, but this obviously requires a stark inconsistency in their beliefs. In particular, (parts of) their belief system will lack justification, since justification is arguably incompatible with stark or deep-running inconsistencies. The cognitive dissonance, we can only imagine, will be hard to bear. Again, this decreases the well-being of the citizens.

What's more, however, it might be inherently instable. There seems to be a human tendency to address and resolve at least the obvious and stark inconsistencies in our beliefs. This does not universally hold for all inconsistencies and all individuals, but it is plausible to assume that a consensus on constitutional essentials will at least be *more* stable if it doesn't require such inconsistencies and instead the citizens are justified in holding their beliefs. This connects to a similar argument by Rawls which we will discuss in the next chapter (section 2.1.2).

Finally, we have to consider the means with which we can bring about such an unjustified agreement or keep it stable in the face of citizens trying to resolve their inconsistencies. These means will, arguably, require some kind of propaganda or indoctrination as well as serious limitations on freedom of speech, press and conscience. And this is by many seen as morally problematic. Again, for liberal democracies the problem seems more pressing than for authoritarian regimes. Thus, if we grant that option 1 is not viable, i.e. a consensus on constitutional essentials is necessary, and still keep aside

option 4 for the moment, then liberal democracies in particular will have a hard time: Either they abolish or prevent a pluralism of worldviews by oppressive means, thereby making a justified consensus on constitutional essentials possible. Or they leave the pluralism as it is, but bring citizens to believe in the consensus on constitutional essentials at the cost of their justification and stabilise these beliefs with oppressive means. Both options in their own way will require oppressive means which are not available to a liberal democracy.

This concludes my discussion of the first three options: First, consensus on constitutional essentials is good, because it promotes stability. Liberal democracies are particularly reliant on this. Second, pluralism is often a fact that cannot be changed (or prevented) without significant moral costs, e.g. the moral costs of forced segregation, indoctrination or rejecting refugees. Again, liberal democracies in particular will have a hard time avoiding pluralism. Third, forcing a combination of pluralism and consensus will lead to serious inconsistencies in (some of the) citizens' beliefs leading to them not being justified. The generation and stabilisation of this state arguably again requires oppressive means with their moral costs. Again, liberal democracies will have particular trouble with this option. You might wonder: Do *only* liberal democracies have problems here? I think that even if a society's constitution is not liberal or not democratic, an overlapping consensus (i.e. joint presence of pluralism, consensus and justification) is still the gold standard for societal stability, since it requires no propaganda apparatus or massive police and military force. But since the problem is particularly pressing for liberal democracies, and I am a liberal democrat, I will henceforth assume that the societies we are talking about are such systems.

So let's turn to option 4. The good news is that pluralism is not always incompatible with a justified consensus on constitutional essentials. What's more, there seem to be liberal democracies in which all three conditions are satisfied, at least to a significant extent. For example, in German society there are several different religions giving (partially) incompatible rules for life as well as many atheists who also disagree on many moral and political questions. Thus, there is a significant pluralism of worldviews. In particular, Germany exhibits a high *moral* diversity when compared to other societies (Osborne and Atari, 2024). Nonetheless, there is a widespread endorsement

of the political system (BPB, 2021). It is, of course, unclear to what extent German citizens are justified in their beliefs. They are certainly not ideal agents without any epistemic deficiencies, but it seems equally clear that it's not the case that the pluralism of worldviews can coexist with the constitutional consenus *only* due to epistemic deficiencies, stabilised by oppressive means. Instead, it seems that there are many worldviews that can, at least in principle, be justifiedly held together with the constitutional essentials that are realised in Germany. Of course, German society is just one such example.

At the same time, it is clear that there are many pluralist societies in which this (justified) consensus is threatened or non-existent. Again, think of the USA, Iran, Israel, Brazil, etc. Thus, there seem to be conditions such that an overlapping consensus is possible and conditions such that this possibility is threatened. As a consequence, the following is a highly relevant question:

> *Under which conditions is an overlapping consensus possible, i.e. a constellation of justified belief systems with a pluralism of worldviews and a consensus on constitutional essentials?*

Answering this question will help determine how viable option 4, i.e. bringing about such conditions, is and whether it is preferable to the first three. In other words, given the considerations about the first three options, answering this question will uncover the conditions under which a democracy can be both stable *and* liberal. In what follows, I lay out which aspect of the question I will be focusing on and my methodology of addressing it.

It is clear that the question is extremely broad. In particular, it has both a descriptive and a normative element, because we want the beliefs to be both existent and justified. On the descriptive side, we can ask about the conditions that need to hold such that beliefs with certain (in part normative) properties exist in a society. This is a question for psychology, cognitive sciences, sociology, and perhaps others. On the normative side, we can ask what conditions need to hold such that the justifiedness criterion permits for a constellation of belief systems to exhibit both pluralism and consensus. This is at its core an epistemological question and it is this aspect that the present thesis focuses upon.

Perhaps an analogy can help see what exactly this normative element is and what it contributes. Suppose you wish to find out under which

conditions people can act morally without acting against their other personal interests. On the normative side of this investigation, you might want to a give a characterisation of 'acting morally' and then contemplate how this characterisation can be in line with or in conflict with acting in accord with other personal interests. This is a necessary first step and gives relevant insights, not least a conceptual toolkit for thinking about these matters. But you will likely also want to empirically investigate the relevant boundary conditions that are given by people's psychology, sociology, etc. After all, it might be the case that something that is theoretically possible will not be in fact realisable due to such boundary conditions. Likewise, my investigation into the possibility of an overlapping consensus will be concerned with the first, normative step: characterise the normative notion (justified consensus among differing worldviews) and develop a first map of how and when it is generally possible. Guided by these insights and the relevant concepts that were developed, we can then investigate how to bring about an overlapping consensus given the actual psychological, sociological, and other empirical boundary conditions.

Now, out of all kinds of conditions that can influence the possibility of an overlapping consensus, I am interested in the dialectic kind. Citizens in a society do not form their beliefs, political or otherwise, in isolation. Instead, they consider the worldviews and arguments of their fellow citizens. At least, they *should* be doing that in order to be justified. In my terminology, every citizen is in a *dialectical situation*. This dialectical situation encompasses at least the views and arguments that are publicly debated in their society, or so I argue later. That means, even if the dialectical situation of a particular citizen encompasses more than that (depending on their circumstances), the dialectical situations of the citizens still have a *common core*, namely the publicly debated worldviews and arguments. Presumably, this common core of the dialectical situations will significantly influence the possibility of an overlapping consensus. For example, the citizens in the above example society of zealots and atheists have a very different common core of publicly debated worldviews than the citizens in, say, German society. To sum up, my focus will be on the following narrower research question:

> *How does the common core of the citizens' dialectical situations, i.e. the publicly debated worldviews and views on constitutional essentials,*

influence the possibility of an overlapping consensus?

What is my methodology of addressing this question? One standard philosophical approach starts by clarifying the involved concepts (in particular: consensus, pluralism and justification) by engaging in *conceptual analysis*. That is, one can address the question by giving (full or partial) definitions of these concepts, e.g. using thought experiments or comparisons with related concepts, and then arguing (perhaps using auxiliary premises) that certain interesting facts about the research question follow from these definitions. This is not the methodology that I will follow. Instead, I follow the methodology of *formal epistemology*. Formal epistemology uses tools from mathematics (including formal logic) to learn about philosophical problems.

In general, this is a four-step process:

1. *Give formal explications of the relevant concepts*: The general idea behind formal philosophy is to represent natural language concepts by mathematical objects.

   > Example: Suppose you want to find out how to rationally change your degree of belief in a certain proposition given some piece of evidence you just discovered. In the first step, a formal epistemologist might represent the natural language concept of 'rational degree of belief' by a mathematical function assigning probabilities to propositions. Also, they might give a general updating rule on how to rationally change these probabilities in the face of evidence. (This is so-called Bayesian epistemology, for an overview see Lin, 2024.)

2. *Model the problem*: Using the formal framework of the explication, the epistemologist models the situation by specifying relevant boundary conditions.

   > Example: Using a representation rule, i.e. a rule on how to translate your degrees of belief to a probability function, you represent prior degrees of belief (i.e. before discovering the evidence) by prior probabilities. (In Bayesian epistemology, this representation rule typically involves betting behaviour.)

3. *Extract formal result*: Using calculations, proofs or (in my case) simulation studies, the epistemologist extracts what the formal apparatus developed in the previous steps says about the question.

   > Example: Using the Bayesian rule for updating probabilities, your new probability for the target proposition is calculated.

4. *Interpret formal result*: Since we are ultimately interested not in mathematical results but in an answer to the philosophical question, in a last step the epistemologist applies the formal result to the original non-mathematical problem in order to get an answer to the original question.

   > Example: Using the representation rule, you can back-translate the formal answer in order to give an answer to your original question, i.e. how to change your degrees of belief in the light of the evidence you discovered.

In essence, the idea behind formal philosophy is to represent the problem by mathematical objects, and then investigate these objects and their proporties in order to learn something about the non-mathematical, real world.

This methodology has advantages and disadvantages. The biggest advantage is its in-built *precision*. Most natural language concepts (including consensus, pluralism and justification) are to some extent vague. This vagueness can sometimes make it hard to give informative answers to a given question. It sets a general limit to how far we can get using the above-mentioned method of conceptual analysis. This is particularly obvious for cases in which trade-offs have to be made. For example, suppose that you have two pieces of evidence, one speaks in favour and the other speaks against a given proposition. What is the overall change in your rational degree of belief in the proposition? Should you assign a lower or a higher degree of belief? Bayesianism can give a definite answer even in cases in which this would be unclear if you only considered the vague natural language concepts involved.

One major disadvantage of formal modeling, especially when a particular framework like Bayesianism is extensively researched, is that one might be so caught up in the intricacies and perhaps internal problems of the formal

framework that the connection to real-world problems is gradually lost. This problem is amplified if the formal framework is so complicated that there is an entry barrier for 'non-formal epistemologists' and, as a result, there is less communication between philosophers with differing methodologies. Formal epistemologists can (and should) try to counter this problem by putting heavy emphasis on trying to make the philosophical foundations and consequences of their work as clear as possible and by connecting these to the respective non-formal philosophic debate. (In this thesis, chapters 2 and 6, respectively, are dedicated to this.)

A few remarks about explications: Explications of a concept, in contrast to definitions, do not aim to perfectly capture every aspect of the natural language concept. Instead, an explication is supposed to make the concept fruitfully precise. Carnap, the inventor of explications, thought that such a fruitful precisification should then *replace* the original vague and perhaps ambiguous natural language concept (Carnap, 1963, p. 3). This is an idea I do not share for the present purposes. Even though I will offer various explications throughout the thesis, in particular, for different kinds of overlapping consensus, none of these should be thought of as a replacement of the natural language concepts. Instead, the picture I embrace is that there are many plausible explications for the natural language concepts that are involved in the research question. We can learn something about the present problem by investigating different such explications and extracting the philosophical consequences of these. If there are some consequences that all explications share, then this gives us confidence that we have found a (partial) answer to the research question. Put differently, an important part of formal modeling involves investigating the robustness of the obtained results. Do the results change significantly if the formal model is varied? If yes, then perhaps the results are mere artifacts of the specific modeling approach and not telling about the philosophical problem. The present thesis makes the first step by giving a handful of similar explications based on a particular formal model and extracting the results they give. In subsequent works, the robustness of these results will have to be investigated.

Now, how exactly do I apply this methodology to the question of the possibility of an overlapping consensus? The core concept that my explications focus on is that of justification. Of course, there is a big epistemolo-

gical discussion about what 'being justified' means. In ethics and political philosophy, many subscribe to the idea that our beliefs in these areas are justified by the *method of reflective equilibrium* (MRE). The idea was first put forth by Goodman (1955) and was later popularised by Rawls (1999). In the next chapter, we will encounter some of its proponents. According to this method, we start from a set of beliefs (the characterisation of this set differs between authors), we then seek a first theory that more or less fits these beliefs and iteratively adjust beliefs and theory to each other until the fit is perfect and the overall belief system forms a coherent whole. Despite its popularity, the specifics of this method are often left unclarified. This is particularly problematic, because these adjustments often involve trade-offs. For example, some theory might be closer to my current beliefs, while some other theory is more elegant and unifying. Which theory should I choose in such cases? Without precise rules, application of this method is at best arbitrary and at worst impossible. This is a prime example of a case in which formal epistemology comes in handy. If we can give an explication of this method, then we have clear rules on how to make such decisions.

Luckily, Claus Beisbart, Gregor Betz and Georg Brun (2021) have recently developed a formal model of the method of reflective equilibrium (henceforth "BBB model"). The model gives a precise algorithm that indicates how exactly to proceed from the starting point (the so-called *initial commitments*) and when to stop, i.e. when the so-called *fixpoint* of an equilibration process is reached and the beliefs are justified. The model is based on the *theory of dialectical structures* by Betz (2021). Accordingly, a dialectical structure (a set of sentences connected by arguments) is the background for any equilibration process, i.e. for any application of the method of reflective equilibrium. An agent occupies an initial position in this structure (accepts certain sentences) and during equilibration changes this position according to the algorithm. One central idea in this thesis is that the above-mentioned dialectical *situation* of a citizen is represented by the dialectical *structure* in which an agent positions themselves during equilibration.

As a consequence, the BBB model of MRE gives an explication for the concept of justification. (This is step 1 in the above four-step sequence: giving explications.) How can I use this explication to study the realisability of an overlapping consensus? In principle, it would be possible to apply

the model to empirical data from an actual society. That is, we might empirically determine the initial commitments and dialectical structures of real citizens and run computer simulations using the algorithm of the BBB model to determine the fixpoints, i.e. the belief systems that are justified for the citizens. We can then investigate whether there is consensus and pluralism in the fixpoints, whether the actual beliefs of the citizens match their respective fixpoints, etc. No doubt we would learn much about the overlapping consensus in this society or the possibility thereof.

Unfortunately, this is currently computationally infeasible, because the structures would be too large to run the simulations in reasonable time. Moreover, we would only learn something about the particular society under investigation and I am more interested in the general rules for overlapping consensus. Thus, I will instead simulate small, artificial, randomly generated societies with agents that have comparatively small dialectical structures. (This is step 2: modelling the problem.) For each such society, I can then investigate whether the justified belief systems of the agents (i.e. the fixpoints) exhibit a pluralism of worldviews and a consensus on constitutional essentials. (For this purpose, I will offer appropriate explications of the latter two concepts.) If there is both pluralism and consensus among the fixpoints, then there is an overlapping consensus. Some of these artificial societies will exhibit an overlapping consensus, others won't. We can then check under which conditions an overlapping consensus is likely and under which conditions it is not. (This is step 3: extracting results.) More precisely, I will generate these artificial societies in a way that let's me isolate how the dialectical structures of the citizens influence the probability of an overlapping consensus.

What do the results about these artificial, randomly generated societies tell us about real-world societies? First, we must check whether the results are robust when the models of the artificial societies are gradually de-idealised such that they are more similar to real-world societies in size and complexity. If a gradual, feasible de-idealisation does not change the results, then that gives us some confidence that the results will also hold for artificial, randomly generated societies of real-world size and complexity (even if we cannot simulate these). Suppose the results are, in fact, robust. Since the simulated societies are *randomly* generated, they are a random

sample of possibility space as a whole. Thus, we may infer that the rules holding for the simulated societies hold for the whole of the possibility space from which they are sampled. This is like inferring that the results of a study about drug efficacy hold for the population from which the participants of the study were sampled (given that the sampling is good). Thus, we have some confidence that we have found general rules that hold for all societies from this possibility space. The real-world societies (or their idealised counterparts) occupy a part of this space.

Thus, in a second step, we can then infer that these general rules will hold for real-world societies as well, as long as there is no defeating evidence. (This is step 4: interpreting results.) To stay in the analogy regarding drug efficacy, this is like the inference from what we found for the whole population to what can be expected in individual cases. This is, in essence, how the results for the artificial, randomly generated societies inform us about the real-world problem. Of course, it is always possible that the more realistic subset of the possibility space as a whole shows a somewhat different behaviour, just like a certain class of individuals might react differently to a drug than the population as a whole. Optimally, in a further step we bring empirical data into the equation. What characterises the part of possibility space occupied by real-world societies? Do the general rules we found hold here as well? To stay in analogy, it is sensible to test drug efficacy in ever smaller subsets of the population: Does the drug work for women in particular, for pregnant women, for pregnant women with certain genetic conditions, etc? But, again, I will start as a first step by trying to find *general* rules for the possibility of an overlapping consensus.

This concludes my introduction to the research question and my methodology of addressing it. The rest of the thesis is structured as follows.

Chapter 2 is concerned with the philosophical foundations for the more formal content of the chapters that follow. It contains all my philosophical commitments and assumptions. A central point of reference will be the works of John Rawls, since he delivered the canonical description of the problem of pluralism for liberal democracies and suggested the idea of an overlapping consensus as the solution. However, the present research is in many ways independent of the specifics of the Rawlsian account and I will be clear on what I share and what I don't share. In fact, chapter 2

quite generally serves to connect the present thesis with related literature concerning the different topics that will come up. Section 2.1 is concerned with the idea of an overlapping consensus. I lay out the basic components, i.e. pluralism, consensus and justification, and highlight once more why an overlapping consensus is the gold standard of stability in liberal democracies when discussing Rawls's argument for this.

In section 2.1.3, I carve out new conceptual territory beyond Rawls by distinguishing different kinds of the notion of an overlapping consensus, in particular, the distinctions of an *actual vs. potential* and *global vs. local* overlapping consensus. These different notions correspond to different stages of an overlapping consensus.

The final stage that we are ultimately interested in is that of an *actual* global overlapping consensus. This kind of overlapping consensus is characterised by the simultaneous satisfaction of the three above conditions: All citizens hold justified belief systems, and the combination of their belief systems exhibits a pluralism of worldviews and a shared view on constitutional essentials. A *potential* global overlapping consensus, on the other hand, does not require that citizens actually hold justified belief systems. Instead it only requires that *if* citizens held belief systems that are justified for them, then these belief systems *would* exhibit a pluralism of worldviews and a shared view on constitutional essentials. Things are a bit more complicated than that, because for any citizen, there might be several belief systems that are justified for them. In every society there is what I call a *space of justified belief systems* containing all constellations of justified belief systems. Some of these might exhibit pluralism and consensus while others might not. I distinguish different kinds of potential global overlapping consensus depending on how many of these combinations exhibit pluralism and consensus. The notion of a potential overlapping consensus is central to this thesis. It marks an intermediary stage that might be transformed into an actual overlapping consensus by assisting citizens in forming justified belief systems. Importantly, since for this thesis I am simulating only artificial societies, there is no relevant use for the notion of an 'actually held belief system'. As a consequence, when analysing the simulation results I will exclusively talk about the potential kind of overlapping consensus.

Additionally, even if there is no *global* overlapping consensus of either

the actual or potential kind, there might still be a *local* overlapping consensus in the sense that there is a group of citizens (forming a *subsociety*) who agree on a shared view on constitutional essentials while holding a pluralism of worldviews. If there is such a local kind of overlapping consensus (of either the actual or potential kind), then it follows that in this subsociety conditions hold such that the pluralism of worldviews does not stand in the way of consensus and justification. As a consequence, it can also mark an intermediary step towards the global kinds of overlapping consensus: Abstractly speaking, we can try to bring it about that the favorable conditions of the subsociety hold in the rest of society as well. In section 6.1 I give a more concrete example of what this can mean.

Section 2.2 is concerned with the relevant notion of justification that is presupposed in the concept of an overlapping consensus. I adopt the idea of Rawls and many others that the method of reflective equilibrium is the correct criterion of justification here. However, there are a lot of conceptual questions to be answered. I discuss these questions and make explicit what assumptions I make and to some extent defend these assumptions. In particular, I commit to the idea that the method of reflective equilibrium is centrally about *coherence* which (at least) involves that one's beliefs can be derived from a systematic theory. I discuss the notion of a dialectical situation, i.e. the views and arguments one needs to consider during equilibration, and propose an alternative to the Rawlsian account: I argue that the citizens' dialectical situations include at least the worldviews and constitutional essentials that are *publicly debated* in their society. Furthermore, I commit to Reconstructionism (agents need not actually apply any specific method of reflective equilibrium), epistemic consequentialism (being in reflective equilibrium is the epistemic goal and being justified means choosing appropriate means for this goal) and a bounded rationality perspective (these means have to be feasible for non-ideal epistemic agents).

Finally, in section 2.2.6, I formulate a research hypothesis about how the common core of the dialectical situations of the citizens (given by public debate) influences the possibility of an overlapping consensus. The focus of this hypothesis will be on the inferential connections between the (publicly debated) worldviews and the conception of constitutional essentials that citizens might agree on. There are three possible connections between

any pair of worldview and conception: Either the worldview supports the conception, or it is neutral about it, or it is incompatible with it. Offhand, support connections seem to be best for an overlapping consensus on the conception. If many of the publicly debated worldviews support a particular conception of constitutional essentials, then an overlapping consensus on that conception seems more likely. Incompatibility, on the other hand, is plausibly going to make it less likely. Neutrality is an unclear candidate. Since the support connection is the only connection of which it is initially plausible to say that it will make an overlapping consensus more likely, I formulate a research hypothesis that states (very roughly): *Most publicly debated worldviews must support a conception of constitutional essentials such that an overlapping consensus on that conception is likely.* Of course, since I have distinguished different kinds of overlapping consensus, this gives us several research hypotheses, one for each kind. The central goal of the simulation study presented in the later chapters is to test these hypotheses. If these hypotheses turn out to hold, then this sets a demanding standard for the realisation of an overlapping consensus. It would be best if many different kinds of worldviews, also neutral or even incompatible ones, can be publicly debated without impinging too heavily on the possibility of an overlapping consensus.

In chapter 3, I apply the BBB model to the philosophical foundations given in the previous chapter by proposing a set of *formal explications* of the different kinds of overlapping consensus. First and foremost, this involves a formal explication of the notion of justification, which I give in sections 3.1–3.3 using the formal BBB model of reflective equilibrium. The model consists of two parts: the so-called achievement function representing the degree to which a belief system is in equilibrium and an algorithm for changing one's belief system in order to optimise this function. The basic idea for explicating justification is this: Whenever a belief system could have been the outcome of applying the algorithm, then that belief system is justified. However, the standard algorithm used in the BBB model is computationally demanding, setting too high an epistemic standard for citizens (violating the bounded rationality condition from the previous chapter) and making simulations of even moderate complexity computationally unfeasible. For these reasons, I advocate for using a different algorithm that optimises in a more step-wise

manner which makes it much more feasible for computers and human brains alike. I argue that the resulting model of reflective equilibrium can thus far be counted as a plausible one.

In section 3.4, I present a measure of consensus (namely: acceptance rate), and three measures of pluralism (namely: entropy, option count and strength of the weak). Each measure of pluralism focuses on a different aspect of the notion, they are normalised in order to give comparable results. Finally, I fuse these ingredients into a set of explications of the different kinds of overlapping consensus. These explications will be the basis for addressing the research question and hypotheses when analysing the results in chapter 5.

Chapter 4 lays out the design of the simulation study. Any kind of formal modelling involves idealisations and my study is no different in this regard. Section 4.1 lays out, explains and motivates these idealisations. The result is a set of conditions for the artificial societies I wish to simulate. The set of societies conforming to these conditions is, thanks to combinatorial explosion, way too large to simulate in its entirety, even when using the more frugal step-wise algorithm. Thus, I need a way of sampling this possibility space and I will *not* do so by simply sampling with a uniform probability distribution. In section 4.2, I present a fair, question-oriented and computationally feasible way of sampling the possibility space. In particular, I propose a method (involving mathematical objects called *multi-sets*) for greatly reducing the complexity of the possibility space without losing information that is significant for the research question. Finally, in section 4.3, I discuss once more how the simulation of the artificially constructed societies tells us something about real-world societies, and what further work needs to be done such that this gap between simulation and real world can be reduced.

Chapter 5 is dedicated to presenting, explaining and interpreting the results of the simulation study. The goal of section 5.1 is to get a grip on how the different connection types (support, incompatibility and neutrality) between worldviews and a given conception of constitutional essentials influence consensus and pluralism in the justified belief systems of the agents. The data is presented by using a somewhat complex but very informative kind of plot, namely ternary heatmaps. I explain how to navigate these heatmaps and discuss the plots for the different measures of consensus and

pluralism. I do not only describe, but also to some depth explain the trends in these plots. The goal of this explanation is a basic plausibility check, i.e. a check whether the simulation study yields results that can be understood without making implausible assumptions. I argue that the study by and large passes this plausibility check. However, it will turn out that in some special cases, the small number of agents in the artificial societies I simulated interferes with the normalisation of the pluralism measures. Whenever this is the case, I supplement with simulations of larger societies.

In section 5.2, I go through the potential kinds of overlapping consensus and analyse what the data says about their probability given different combinations of supportive, incompatible and neutral worldviews. This analysis is somewhat complex, but the general upshot is this: A global overlapping consensus on a conception of constitutional essentials is only probable if most worldviews support the conception. That is, the research hypothesis about global overlapping consensus fits with the data. However, a local overlapping consensus is probable even if it is not the case that most worldviews support the conception. In fact, even if there are only neutral worldviews in the dialectical situations of the citizens, a local overlapping consensus is still probable. The hypothesis about local overlapping consensus is falsified. Section 5.3 gives an overview of the study results and their verdict about the research hypotheses.

Chapter 6 concludes and presents ideas for follow-up studies. In particular, I give a comprehensive and somewhat detailed overview of the thesis. I recall the most important philosophical as well as modelling assumptions. I condense the diverse and relatively complicated results from chapter 5 into a shorter and more easily accessible upshot. Also, I draw something like a main conclusion from the thesis: There is hope for an overlapping consensus, i.e. hope for avoiding the undesireable options 1–3 above. The reason for this hope is that even if all publicly debated worldviews are neutral, there may still be a local overlapping consensus. Of course, what we are ultimately interested in is a global overlapping consensus. But, as I have mentioned above, we might be able to turn a local overlapping consensus into a global one. I give a suggestion as to how this might be done, a suggestion that is testable in future simulation studies. In any case, there is still much work to be done. In particular, the robustness of the present results needs to be

investigated. I distinguish three levels of studying robustness: varying the study design, varying the model of reflective equilibrium, and considering altogether different ways of explicating justification. I make specific suggestions for each level. Thus, the present thesis does not pretend to give final answers itself. Instead, it aims to carve out a research programme that has the potential to yield results that are both robust and of practical relevance.

Chapter 2

Overlapping consensus

The first goal of this chapter is to lay the philosophical foundations for the ones to follow. In fact, I think this distinction between the first, philosophical chapter and the subsequent, formal chapters is pretty strict: I have attempted to put all important philosophical assumptions and arguments into this first chapter such that the philosophical foundation of the formal work is as transparent as possible. The second goal of this chapter is to situate the present thesis in the wider research context: How does it relate to other research on political liberalism? How does it relate to other research in epistemology? There is no dedicated section for this, instead, I will do so along the way. The third goal is to give a precise and relevant research question as well as testable hypotheses.

In the first section 2.1 I present Rawls's account of overlapping consensus, the general idea of which I adopt. Additionally, I develop a bunch of interesting distinctions between different kinds of overlapping consensus, most importantly: actual vs. potential and global vs. local. The second section 2.2 is about the notion of justification that I take to be most relevant here. There are a number of philosophical decisions to make which to some extent guide the development of the formal notions presented later and, more importantly, fix their interpretation. With these commitments, I present and motivate a research question and hypotheses that will be tested in the later chapters. The last section 2.3 summarises my philosophical commitments and fuses them into a definition of justification. This definition will be the basis for giving explications of the different kinds of overlapping consensus in the next chapter.

2.1 Overlapping Consensus

I begin by laying out the general Rawlsian idea of an overlapping consensus, to which I am committed, and highlight which aspects of it I am not committed to.

I should note from the start that Rawls's political philosophy is both broad and deep. Over the years he has worked it out in great detail and sometimes significantly changed his position. I make no pretense of covering the many issues that he has worked on. For an excellent outline of his latest view, see (Wenar, 2021). The idea of an overlapping consensus plays an important role in Rawls's philosophical system. However, I do not want to investigate the realisability of an overlapping consensus because I am convinced of his system as a whole. Instead, I think that the general idea of an overlapping consensus account of societal stability is plausible in itself, as I have argued in the introduction and will highlight again in section 2.1.2. As a consequence, I will *cherry-pick* this idea from Rawls. Importantly, I am only partially concerned with the intricate role that the concept of an overlapping consensus plays in his overall system. In particular, I won't discuss the role it plays in what Rawls calls a 'public justification of a political conception of justice' (see section 2.2.2), even though this is no doubt one of his prime goals. Also, I won't argue in detail that the explication of overlapping consensus that I offer in the next chapter can fulfill all of Rawls's intended purposes. Instead, my discussion of related Rawlsian ideas will mainly serve to avoid misunderstandings and clarify the philosophical foundations of the later, more technical chapters of this thesis. Nonetheless, I do think that the present research is of interest to many political philosophers who consider themselves liberal democrats, including party line Rawlsians. Throughout this chapter, I will highlight in how far my assumptions are compatible with different views on the matters that I touch upon. Often I am able to take a rather non-committal stance, which I take to be a good thing, because it makes my research relevant for philosophers of different camps.

2.1.1 Rawlsian overlapping consensus

In his 1971 *A Theory of Justice* (abbr. 'TJ', references to the revised edition (1999)), John Rawls formulates and aims to justify *justice as fairness* (JF), a theory of justice concerning the basic structure of society. The theory consists of two principles of justice, lexically ordered in priority. Often labelled a version of 'egalitarian liberalism', JF is an alternative to the utilitarian paradigm. Since I am not concerned with JF's content in detail (see below), I won't review it here. In the follow-up monograph *Political Liberalism* (abbr. 'PL', references to the expanded edition (2005)), Rawls addresses what he considers fundamental shortcomings in TJ. The most serious problem is that his account of societal stability in TJ, the so-called *well-ordered society*, is unrealistic: "An essential feature of a well-ordered society associated with justice as fairness is that all its citizens endorse this conception on the basis of what I now call a comprehensive philosophical doctrine", even though "a plurality of reasonable yet incompatible comprehensive doctrines is the normal result of the exercise of human reason within the framework of the free institutions of a constitutional democratic regime" (PL xvi). Thus, the realisation of JF, providing this framework, under normal circumstances leads to that society being pluralist and not well-ordered in the sense put forward in TJ.

The first step towards a solution to this problem is to emphasise the distinction between the moral in general and the purely political:

> "In my summary of the aims of *Theory*, the social contract tradition is seen as part of moral philosophy and no distinction is drawn between moral and political philosophy. In *Theory*, a moral doctrine of justice general in scope is not distinguished from a strictly political conception of justice. Nothing is made of the contrast between comprehensive philosophical and moral doctrines and conceptions limited to the domain of the political. In the lectures of this volume, however, these distinctions and related ideas are fundamental." (PL xv)

Even though Rawls clarified earlier (1985) that JF must be understood as a purely political conception of justice, only in PL did he fully acknowledge that JF was in TJ presented as rooted or grounded in a comprehensive doc-

trine that citizens might reasonably reject. His central goal in PL, then, is to *reformulate* JF as a political conception of justice (cf. xli). This means, most importantly, that JF is presented as a freestanding view, i.e. formulated in doctrine-neutral terms, such that it can fit as a module into various reasonable comprehensive doctrines (cf. PL 12). Rawls describes the challenge as follows:

> "[T]he problem of political liberalism is: How is it possible that there may exist over time a stable and just society of free and equal citizens profoundly divided by reasonable though incompatible religious, philosophical, and moral doctrines? Put another way: How is it possible that deeply opposed though reasonable comprehensive doctrines may live together and all affirm the political conception of a constitutional regime? What is the structure and content of a political conception that can gain the support of such an overlapping consensus?" (PL xviii)

Rawls describes that the new account of stability does not build on the idea that the theory of justice regulating society is affirmed by all citizens on the basis of the same comprehensive doctrine. Instead, it builds on the idea of an *overlapping consensus*: Various comprehensive doctrines 'live together' despite their incompatibility and all 'affirm' the same political conception, i.e. they overlap on it. (For a very detailed and influential reconstruction of Rawls's political turn, see Weithman (2010). I agree with the general gist of Weithman's analysis.)

I am convinced by Rawls's proposal. Despite some criticism that it has met (e.g. Hampton, 1989; Raz, 1990), I will not defend the general idea in this thesis. Instead, it should be viewed as a fundamental presupposition. In particular, both the distinction between the moral in general and the purely political as well as the concept of an overlapping consensus are the main ideas I draw from Rawls's political liberalism. However, I am not committed to most (and skeptical of some) of the Rawlsian details. He is mainly concerned with developing and reflecting on a new version of JF, a freestanding liberal political conception of justice, including Rawls's ideas of autonomy, political constructivism, public reason, and so forth. These ideas will not concern me in the following chapters and I am to no extent committed to them.

What I am committed to is the idea that pluralist societies can be stable if the worldviews (or 'comprehensive doctrines' in Rawls's terms) overlap sufficiently when it comes to basic political questions.

There is a further complication in Rawls's work that I won't take into account in what's to come. At some points he speculates that the focus of an overlapping consensus, i.e. the common ground of the doctrines, may not necessarily be a single political conception of justice but instead a "class of liberal conceptions that vary within a certain more or less narrow range" (PL 164). However, most often Rawls takes the focus to be a single conception and he does not really work out this idea of 'justice pluralism' in more detail (though Weithman, 2023, picks up the slack). For this reason, and for the sake of simplicity, I will not consider this idea. However, it may be that it can rather straightforwardly be connected with the present work. In fact, if a class of political conceptions can simply be represented by a disjunction of political conceptions, then the results of the present investigation also holds for such 'classes' of conceptions. But it's less clear how to investigate the internal complexities of such a class and their influence on the possibility of an overlapping consensus. In any case, I will henceforth continue speaking of overlapping consensus on a single political conception of justice.

You might ask whether I can really be non-committal about the content of the entities involved in an overlapping consensus (doctrines, conceptions) to the point that the focus of an overlapping consensus may even be a disjunction of doctrines and not a single one. But this is a general feature of the present thesis. As we will see later, I adopt a purely structural perspective, focusing solely on the *inferential relations* between the worldviews and the focus of the overlapping consensus. This is not unlike an investigation into the logics of a subject matter. As a consequence, the results of the present thesis are compatible with many different views on what the difference is between the moral in general and the purely political, on what constitutes a worldview or comprehensive doctrine, on how to make a conception of justice freestanding, etc. (This point will become much clearer when I return to this issue in section 4.3.) I take this neutrality to be a strong suit. Anyone who finds the general picture of an overlapping consensus appealing, for whatever reasons, might be interested in the present research. Nonetheless, I do adopt the Rawlsian terminology of 'comprehensive doctrines' to denote

what I called worldviews in the introduction and 'political conceptions' to denote what I called (conceptions of) constitutional essentials. This should never be understood as committing to the specific Rawlsian versions of these notions.

Even though Rawls is mostly concerned with the *content* of a freestanding version of JF, he does say something about the purely structural perspective, i.e. the inferential relations between comprehensive doctrines and political conception. Even though I will later challenge his view on this matter, for the time being it nicely illustrates what an overlapping consensus can be and how it contributes to societal stability.

Offhand, there seem to be three possibilities for any pair of comprehensive doctrine and political conception:

- The comprehensive doctrine *supports* the political conception.
- The comprehensive doctrine is *incompatible* with the political conception.
- The comprehensive doctrine is *neutral* about the political conception, i.e. is neither in support of nor incompatible with it.

It is clear that according to Rawls a comprehensive doctrine that is incompatible with a political conception cannot be part of an overlapping consensus on that conception. The whole point of reformulating JF is to make it compatible with other comprehensive doctrines than the one put forward in TJ.

In most relevant passages on the structure of overlapping consensus, Rawls seems to think of comprehensive doctrines as *supporting* the political conception, though his terminology is not stable. (For example, in the above quote he uses the terms 'affirm' and 'support', in other passages the terms 'rest on', 'endorse', and others.) Two such passages are particularly relevant for our purposes. The first one gives a model case which nicely illustrates the core idea of an overlapping consensus. The second one explains how an overlapping consensus contributes to societal stability. Let's start with the model case and discuss the second passage in the next section.

In PL (Lecture IV, § 6 "Conception and Doctrines: How related?"), Rawls gives a model case with three comprehensive doctrines, each of them supporting the same liberal political conception of justice in their own way.

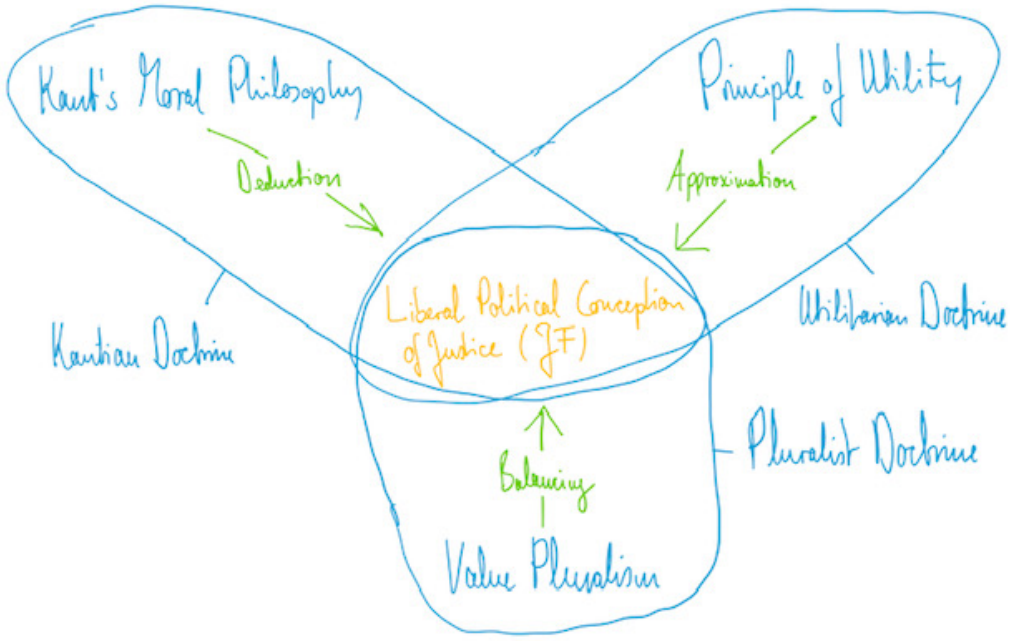

Figure 2.1: Graphic representation of an overlapping consensus. It illustrates the model case of an overlapping consensus given by Rawls (PL § 6).

First, he supposes that Kant's moral philosophy with its ideal of autonomy might *deductively imply* justice as fairness, even though he admits that the argument will be hard to set out rigorously. Second, he speculates that some version of classical utilitarianism might support a liberal conception of justice as an *approximation*: Limits on knowledge and the complexity of rules may lead a utilitarian to accept a liberal conception as a "satisfactory [...] approximation to what the principle of utility, all things tallied up, would require". Third, Rawls imagines "a pluralist account of the realms of values that include[s] the political conception as the part covering political values". This comprehensive view balances these values against each other such that the political values usually outweigh others. Thus, this view supports the political conception by *balancing*. This model case nicely illustrates an overlapping consensus where all comprehensive doctrines support the political conception. Figure 2.1 is a representation of this overlapping consensus, making the metaphorical term 'overlapping' graphically explicit.

2.1.2 Moral justification, stability and reasonability

The second passage I wish to discuss is supposed to answer an objection to the idea of such a consensus (PL 145ff). The objection complains that an overlapping consensus is at bottom a "mere modus vivendi" and embracing it means abandoning "the hope of political community". Rawls responds by confirming that the hope for "a political society united in affirming the same comprehensive doctrine" should indeed be abandoned, but he rejects that an overlapping consensus is a mere modus vivendi (PL 146f.): A modus vivendi is an agreement made by parties that have conflicting interests but no means to overpower the others. The agreement is preferable for each party when compared to an unresolved power struggle, and this fact everyone knows, so everyone honors the agreement. Importantly, should the power balance shift, and one party gain the option to overpower the others, then it will do so. As Rawls puts it, "social unity is only apparent, as its stability is contingent on circumstances remaining such as not to upset the fortunate convergence of interests" (PL 147).

An overlapping consensus, on the other hand, is not an agreement that is made because it is preferable to power struggle. Instead, it is affirmed on moral grounds:

> "All those who affirm the political conception start from within their own comprehensive view and draw on the religious, philosophical, and moral grounds it provides. The fact that people affirm the same political conception on those grounds does not make their affirming it any less religious, philosophical, or moral, as the case may be, since the grounds sincerely held determine the nature of their affirmation. [...] This means that those who affirm the various views supporting the political conception will not withdraw their support of it should the relative strength of their view in society increase and eventually become dominant. [...] This feature of stability highlights a basic contrast between an overlapping consensus and a modus vivendi" (PL 147f.)

First, this quote shows again that Rawls thinks of the comprehensive views as *supporting* the political conception. Citizens 'start from within' their view and affirm the political conception 'on its grounds'. Second, he thinks that

this support explains how the overlapping consensus contributes to societal stability in a way that a modus vivendi cannot: It does not depend on a power balance and is, therefore, more stable than a modus vivendi which is itself, of course, more stable than a power struggle (see also Rawls, 1985, p. 250).

This passage contains a centrally important part of the concept of an overlapping consensus. It's the idea that in an overlapping consensus, acceptance of the political conception is for each citizen *morally justified*. I share this idea with Rawls. In fact, it is one of the most fundamental philosophical assumptions underlying this thesis, as I laid out in the introduction. Of course, it is conceivable that a society of citizens with morally unjustified beliefs *resembles* an overlapping consensus (as, for example, a modus vivendi does). But unless citizens are morally justified in holding their beliefs, particularly their beliefs about constitutional essentials, any such constellation of beliefs will not qualify as an overlapping consensus.

This moral justification of an overlapping consensus contrasts with the pragmatic justification of a modus vivendi. And this contrast, according to Rawls, explains how an overlapping consensus contributes to societal stability. A pragmatic justification depends on a constellation of interests that may change. A moral justification, on the other hand, is independent of these interests. And this makes societies with an overlapping consensus particularly stable: Rawls often uses the term "stability for the right reasons". I agree with Rawls that an overlapping consensus is more stable than a mere modus vivendi. This, together with the argument I gave in the introduction, is the reason why I think that an overlapping consensus should be of central interest when thinking about stability in pluralist societies. In a sense, it is the *gold standard* of stability in liberal democracies, thus, we should investigate its realisability. Of course, if it turns out to be unattainable, then we may rest content with a surrogate like a modus vivendi.

Two remarks: First, I already said that I will challenge Rawls's assumption that comprehensive doctrines need to support the political conception for an overlapping consensus to be possible. However, this challenge is not directed at the idea that citizens need to be morally justified in order for their beliefs to form an overlapping consensus. Instead, it is directed at the idea that a citizen's comprehensive doctrine needs to support the political

conception for them to be morally justified on affirming the conception.

Second, it follows from Rawls's considerations that in an overlapping consensus the political conception is among the moral beliefs of the citizens. That is, moral justification here just means (epistemic) justification of moral beliefs, while pragmatic justification means (epistemic) justification of beliefs about what is prudent or instrumentally rational. This idea, i.e. that the political conception is a moral conception, is highlighted regularly by Rawls throughout PL and I am committed to this idea (see also section 2.1.3). For critical voices regarding the role of moral considerations in political theory, see Williams (2005, ch. 1) or Geuss (2008).

The role of coercive power for stability

In reaction to the abovementioned analysis of the Rawlsian political turn by Weithman (2010), Klosko (2015) criticises the Rawlsian account of stability, or Weithman's reconstruction thereof. In particular, he criticises the sharp distinction between so-called 'imposed stability' and 'inherent stability' (or 'stability for the right reasons') (Weithman, 2010, §II.1; PL xlii). Obviously, the account of stability that is based on the idea of an overlapping consensus is supposed to be one of *inherent* stability: Citizens accept political decisions and conform to them, because they (justifiedly) accept the political conception of justice. The stability is not *imposed* in the sense that citizens are forced to accept and conform to the political decisions, because the state is willing to use coercive power if they don't.

As I have said in the introduction, an overlapping consensus is the gold standard of stability *particularly* for liberal democracies, because they have less resources to use coercive power than authoritarian regimes. Klosko argues, however, that the stability in real liberal democracies cannot easily be classified as either clearly inherent or clearly imposed. Instead, it is the interplay of coercive power and the citizens' justified acceptance of the constitution that explains their stability (Klosko, 2015, p. 243).

What does this mean for the present project? Do I have to rebut Klosko's arguments, because they question its relevance? I don't think that I have to do so and I wouldn't want to either, because I whole-heartedly agree with him on this point: Stability is neither fully inherent nor fully imposed but

some kind of mixture of the two. Interestingly, Hampton (1989, p. 799ff) makes a point similar to Klosko's and argues that Rawls need not and likely does not disagree.

Thus, given that stability is always a mixture of inherent and imposed stability, my project is supposed to contribute to the inherent part. In particular, I wish to investigate how much inherent stability is possible under which conditions. For example, if we find out that an overlapping consensus is only possible given very unrealistic conditions, or perhaps only a weak kind of overlapping consensus is possible given realistic conditions, then this might make it necessary to use a heavier dose of imposed stability in the mixture. (In the next section, I discuss different kinds of overlapping consensus.) Thus, my project might also be seen to contribute to the following question: How much coercive power is necessary to maintain stability under such-and-such conditions?

Reasonability of citizens and doctrines

This point about stability connects to the *reasonability assumption* Rawls makes about citizens and the comprehensive doctrines they hold. Citizens are reasonable only when "they are ready to propose principles and standards as fair terms of cooperation and to abide by them willingly, given the assurance that others will likewise do" and "those norms they view as reasonable for everyone to accept and therefore as justifiable to them; and they are ready to discuss the fair terms that others propose" (PL 49). Additionally, they accept the so-called burdens of judgment showing that disagreement about comprehensive doctrines is the natural outcome of a liberal democratic regime and not "rooted solely in ignorance and perversity, or else in the rivalries for power, status or economic gain" (PL 58). As a consequence, "reasonable persons see that the burdens of judgment set limits on what can be reasonably justified to others, and so they endorse some form of liberty of conscience and freedom of thought" (PL 60). In essence, reasonable citizens are tolerant and abide by the law willingly given the assurance that others will. As a consequence, in societies with reasonable citizens, imposed stability in the above sense is unnecessary, at least to the extent that such assurance can be given without enforcement by penal law (see my discussion of the

assurance problem in section 2.2.2).

Moreover, reasonable citizens endorse only reasonable comprehensive doctrines. Such a doctrine "covers the major religious, philosophical, and moral aspects of human life in a more or less consistent and coherent manner" and "although stable over time, and not subject to sudden and unexplained changes, it tends to evolve slowly in the light of what, from its point of view, it sees as good and sufficient reasons" (PL 59). Rawls does not specify the content of such doctrines, besides requiring that "a reasonable comprehensive doctrine does not reject the essentials of a democratic regime" (PL xvi). Sometimes he sounds as if reasonable doctrines do not only not reject such essentials, but accept them. For example, he states that "simple pluralism moves toward reasonable pluralism" when citizens' doctrines shift such that they "accept the principles of a liberal constitution" (PL 163f). In essence, reasonable comprehensive doctrines are coherent and change only due to good reasons. (This might best be understood as requiring that such doctrines or the citizens holding them are in *reflective equilibrium*, see section 2.2.) Moreover, reasonable doctrines are in support of a liberal constitution (or at least don't reject it).

These reasonability assumptions, i.e. that citizens are reasonable and endorse only reasonable comprehensive doctrines, are obviously far-reaching. In fact, I think that these two assumptions alone get Rawls halfway towards an overlapping consensus. Perhaps they don't get him all the way there, because there may still be reasonable disagreement regarding the essentials of a liberal democratic regime, but they make sure that citizens are liberal democrats or hold doctrines that support liberal democracy (or are at least compatible with liberal democracy).

Given that Rawls wants to do ideal theory, these idealising assumptions are sensible (PL 55). In fact, ideal theory consists precisely in making such assumptions. According to Rawls's strategy, once it is understood how society can work under such ideal conditions, one can construct a non-ideal theory aimed at realising the ideal (cf. Wenar, 2021, § 2.3). (For criticisms of ideal-theoretic approaches see (Mills, 2005; Stahl, 2022), for an overview of the debate see (Valentini, 2012).) I will not make Rawls's idealising assumptions. In particular, I will not from the outset assume that citizens are tolerant or that comprehensive doctrines are in support of (or compat-

ible with) liberal democracy. Instead, it will at most be the *outcome* of the present investigation that *under certain conditions* citizens are reasonable in the Rawlsian sense and hold comprehensive doctrines that are reasonable in the Rawlsian sense. But even if so, my goal is not to show that under certain conditions the Rawlsian ideal can be realised (even though this might in fact be the outcome). The goal of this research is to investigate the conditions under which an overlapping consensus is realisable. If it turns out that an overlapping consensus is realisable even if some citizens hold doctrines that are incompatible with liberal democracy, then that's fine with me even if such a society does not realise the Rawlsian ideal.

Let's recap the most important takeaways of this and the last section:

- There is an overlapping consensus in a pluralist society only if the different comprehensive doctrines in a society overlap on a shared political conception of justice.
- Citizens in an overlapping consensus are morally justified in endorsing the political conception.
- An overlapping consensus is the gold standard for stability in pluralist societies.
- Even if, as Klosko (2015) maintains and I agree, stability in liberal democracies is *de facto* also partly achieved by the use of coercive power, it is still best if stability is as much as possible secured by the citizens' justified allegiance to the constitution.
- I will not from the outset assume that citizens or the doctrines held by them are reasonable in the demanding Rawlsian sense. Instead, this will at most be the outcome of the present research.

2.1.3 Different kinds of overlapping consensus

Weithman's analysis of Rawls's political turn (as primarily being concerned with stability) was generally well received among political liberalists (e.g. Neufeld, 2011). Nonetheless, there has been little concern with how to realise an overlapping consensus (though see section 2.2.6). Instead, philosophers usually theorise about societies in which an overlapping consensus is already

established. One example of this is the recent debate about which view on public reason is better suited to deal with what Wong and Li (2023) called the *assurance problem*. In section 2.2.2 I turn to this debate in more detail.

The present thesis is concerned with the prior question: the realisability an overlapping consensus. I think that for this purpose it is useful to expand our conceptual toolkit.

In this section, I present definitions for different kinds of overlapping consensus. These definitions are supposed to give a more precise view on the concept of overlapping consensus and, importantly, its preliminary stages. They will play an important part in generating research hypotheses in section 2.2.5 and discussing the results of the simulation study in chapter 5. However, these distinctions cannot be found in Rawls's PL or the related literature, they are my own conceptual contribution. I should note from the start that I will use a semi-formal language to define these concepts. I don't think that this is strictly speaking necessary, but the concepts are expressed and understood more easily and more precisely in this way. Also, the formal explications presented in later chapters will straightforwardly connect to the notions presented here.

A central aspect of an overlapping consensus is that citizens are morally justified in endorsing the shared political conception of justice. The next section of this chapter will be about characterising the relevant kind of justification in detail: justification by the method of reflective equilibrium. One important feature of this kind of justification will turn out to be its *holism*. That is, a particular belief is justified iff it is part of a justified belief system. (A version of this is the famous Duhem-Quine thesis (Quine, 1951). For a recent explicit defense of holism, see Elgin (2005).)

As a consequence, a citizen is justified in endorsing the political conception iff the political conception is part of a system of moral beliefs that is justified for that citizen. This system of moral beliefs encompasses both the moral in general as well as the purely political. Thus, there is an overlapping consensus only if every citizen holds a justified system of moral beliefs. (In what follows, I will often just talk of belief systems, but mean moral belief systems. Likewise, I talk of justification, but mean moral justification.) In section 2.2.5 I discuss the idea that holism about justification holds not only for the moral belief system of an agent, but for their *overall* belief system. For

now, I ignore this complication.

Let $A = \{a_1, ..., a_n\}$ be a set of agents living together in a society. These agents are the citizens of that society. Let B be the set of all possible (moral) belief systems. These belief systems can be thought of as sets of believed sentences. In section 3.1, this notion will be made formally precise. Let $J_i \subset B$ denote the finite set of belief systems that are (morally) justified for agent a_i. This set might contain just one belief system or several. Let $b^{a_i} \in B$ denote the belief system actually held by agent a_i. Let PC be a political conception.

Definition 1 (Actual Global Overlapping Consensus)**.** There is an *actual global overlapping consensus on PC* iff

1. $b^{a_i} \in J_i$ for all $i = 1, ..., n$; and
2. the tuple $(b^{a_1}, ..., b^{a_n})$ exhibits a pluralism of comprehensive doctrines and a consensus on PC.

This definition of an overlapping consensus is the most straightforward one. Condition 2 ensures that there is overlap on the political conception even though the agents endorse a pluralism of comprehensive doctrines. Condition 1 ensures that the belief systems of the agents are justified for them, thus, every agent is justified in endorsing the political conception.

You might wonder: What does it take for a tuple of belief systems to exhibit a pluralism of comprehensive doctrines and a consensus on PC? In section 3.4, I present a formal account of pluralism and consensus. For now, let's suppose that every political conception and every comprehensive doctrine is represented by a respective set of sentences that can be contained or not contained in a belief system. There is a consensus on PC in a tuple of belief systems if all or most of them contain PC. There is a pluralism of comprehensive doctrines in a tuple of belief systems if there is not a particular comprehensive doctrine that is contained in all or most of them. Instead, some belief systems contain one comprehensive doctrine, some contain another, and some contain a third or fourth, etc.

Now, it is clear that sometimes people do not hold justified belief systems. Even so, we might be interested in whether there is a *potential* overlapping consensus even if there is no *actual* overlapping consensus. The modal 'potential' is here to be interpreted as: If each citizen *held* a belief system that is justified for them, then there *would* be an overlapping consensus.

If for every agent there is exactly one belief system that is justified for them (i.e. their set of justified belief systems is a singleton), then the answer is straightforward: There is a potential overlapping consensus iff the tuple of these justified belief systems exhibits a pluralism of comprehensive doctrines and a consensus on *PC*. However, the explication of justification offered in chapter 3 allows for several justified belief systems per agent. In section 3.3, I briefly discuss how this feature relates to the discussion about epistemic permissiveness vs. evidential uniqueness in epistemology (White, 2005; Briesen, 2017). For now, let's keep it general and allow for the possibility that there is more than one justified belief system per agent.

In this case, there is not just one tuple of justified belief systems, but many. For example, suppose there is a society of three agents a_1, a_2, a_3 with $J_1 = \{b_1, b_2\}$, $J_2 = \{b_3\}$, $J_3 = \{b_4, b_5\}$. Then there is not just one tuple of justified belief systems, but $2 \cdot 1 \cdot 2 = 4$: (b_1, b_3, b_4), (b_2, b_3, b_4), (b_1, b_3, b_5), (b_2, b_3, b_5). Every agent has a fixed position in these tuples and this position can be filled by any belief system that is justified for that agent. For n agents, the set of these tuples is an n-dimensional *space of justified belief systems*, denoted by the Cartesian product $J_1 \times ... \times J_n$ with $|J_1 \times ... \times J_n| = \prod_i |J_i|$. These tuples of justified belief systems are not unlike points in 3D space or 4D spacetime.

The question is: Which of these tuples of justified belief systems needs to exhibit a pluralism of comprehensive doctrines and a consensus on *PC* for there to be a potential overlapping consensus? Without further information about the agents, none of these tuples is privileged over the others. I suggest to differentiate different senses of a potential overlapping consensus:

Definition 2 (Potential Global Overlapping Consensus)**.** There is a *potential global overlapping consensus on PC*

- *in the strong sense* iff all tuples from $J_1 \times ... \times J_n$ exhibit a pluralism of comprehensive doctrines and a consensus on *PC*.
- *in the weak sense* iff there is at least one tuple from $J_1 \times ... \times J_n$ that exhibits a pluralism of comprehensive doctrines and a consensus on *PC*.
- *of grade r* iff a proportion $r \in [0, 1]$ of all tuples from $J_1 \times ... \times J_n$ exhibits a pluralism of comprehensive doctrines and a consensus on *PC*.

Obviously, the sense 'of grade r' is a generalisation of the other two senses, i.e. they can be reduced to it: There is a potential global overlapping consensus in the strong sense iff there is a potential global overlapping consensus of grade 1. There is a potential global overlapping consensus in the weak sense iff there is a potential global overlapping consensus of grade $r > 0$. Also note that if there is only one justified belief system per agent then the different senses are equivalent: There is a potential global overlapping consensus in the weak sense iff there is one in the strong sense iff there is one of grade $r > 0$ (equivalently: $r = 1$). That is, in this case there is no reason to make these distinctions.

The different senses of a potential overlapping consensus can be interpreted as:

- Strong sense: If each citizen held a belief system that is justified for them, then it would be guaranteed that there is an overlapping consensus.
- Weak sense: If each citizen held a belief system that is justified for them, then there would a chance, however slight, that there is an overlapping consensus.
- Grade r: If each citizen held a belief system that is justified for them, then there would a probability of r that there is an overlapping consensus (given we know nothing else about the agents, etc).

Given these interpretations, you might ask: What is the relevance of this potential overlapping consensus in whatever sense? After all, we're interested in the real deal, the *actual* overlapping consensus. The answer to this question is that there are different reasons why there is no actual overlapping consensus and depending on this reason we have different options to bring it about. If there is a potential overlapping consensus, especially one of high grade or even in the strong sense, then we might try to bring about an actual overlapping consensus by incentivising the citizens to change their belief system into one that is justified for them. If, on the other hand, there is not even a potential overlapping consensus, then we might have to take a different route, see below. In a nutshell, the existence of a potential overlapping consensus gives us a hint as to what needs to be done for an actual overlapping consensus.

Another reason why investigation into the existence of a potential overlapping consensus is worthwhile is the following: Suppose there is an actual overlapping consensus. As a consequence, there is also a potential overlapping consensus at least in the weak sense, i.e. of grade $r > 0$. However, if this potential overlapping consensus turns out to be of a low grade, e.g. there is only one tuple in $J_1 \times ... \times J_n$ exhibiting a pluralism of comprehensive doctrines and a consensus on *PC*, then this threatens stability in this society. This is because citizens might change their belief systems and thereby loosing the actual overlapping consensus *without ceasing to be justified*. Thus, even if there is an actual overlapping consensus, this actual overlapping consensus is threatened if the potential overlapping consensus in this society is of low grade.

Let's turn to yet another kind of overlapping consensus. As you will have noticed, definitions 1 and 2 mention the term 'global' and I have thus far cheekily avoided to explain it. The term here means simply 'society-wide' and is plainly what we are after when it comes to overlapping consensus. Nonetheless, it might still be interesting to look into overlapping consensus that are *not* society-wide, i.e. into *local* overlapping consensus. Suppose: There is a society without a global overlapping consensus, neither actual nor potential in whatever sense. The problem is that the tuples of justified belief systems do not exhibit a consensus on *PC*. However, suppose there is at least one tuple such that there is a part of this society, a *subsociety* if you will, that exhibits pluralism and consensus on *PC* in this tuple. Then this shows that there is a combination of comprehensive doctrines and political conception such that a pluralism of these doctrines does not itself stand in the way of justified consensus on the political conception. This, in turn, means that not all hope is lost and we might try to bring about a potential *global* overlapping consensus in at least the weak sense by, abstractly speaking, identifying the relevant circumstances that lead to the local overlapping consensus in the respective part of society and try to bring it about that these circumstances hold on the rest of society as well. In chapter 6, it will become clearer what 'changing the circumstances' can mean in this situation.

Let's give a definition for this local overlapping consensus on *PC*. Similarly to definitions 1 and 2, the locus of pluralism and consensus is tuples of belief systems. However, not all belief systems in a tuple need to agree

on *PC*, since the overlapping consensus is supposed to be local. Instead, we only consider the belief systems that accept *PC* and check whether *these* exhibit a pluralism of comprehensive doctrines. If they do, then these belief systems exhibit a both a consensus on *PC* (by definition) and a pluralism of comprehensive doctrines. Let's define this more rigorously:

Definition 3 (Pluralism in *PC*-subsociety)**.** Let $(b_1, ..., b_n) \in J_1 \times ... \times J_n$. Let $I_{PC} := \{i \in \{1, \ldots, n\} : b_i \text{ accepts } PC\}$ with cardinality $m := |I_{PC}| \leq n$ and elements $s_1 < ... < s_m$. The tuple $(b_{s_1}, ..., b_{s_m})$ is called a *subtuple* of $(b_1, ..., b_n)$. More precisely, it is called the *PC-subtuple* of $(b_1, ..., b_n)$. The agents $a_{s_1}, ..., a_{s_m}$ are called the *PC-subsociety* of $(b_1, ..., b_n)$. The tuple $(b_1, ..., b_n)$ is defined to exhibit a *pluralism of comprehensive doctrines in its PC-subsociety* iff $(b_{s_1}, ..., b_{s_m})$ exhibits a pluralism of comprehensive doctrines.

Basically, this definition constructs a new, smaller tuple from a given one by cutting out all belief systems that do not accept *PC*. The remaining ones, in their original order, form the new tuple. This new tuple is called a subtuple of the original one. If this *PC*-subtuple exhibits a pluralism of doctrines, then the original one exhibits a pluralism of doctrines in its *PC*-subsociety.

We have now defined what it means for a tuple to exhibit a pluralism of comprehensive doctrines in the *PC*-subsociety. Using this definition, we can define the notion of a potential local overlapping consensus:

Definition 4 (Potential Local Overlapping Consensus)**.** There is a *potential local overlapping consensus on PC*

- *in the weak sense* iff there is at least one tuple from $J_1 \times ... \times J_n$ that exhibits a pluralism of comprehensive doctrines in its *PC*-subsociety.
- *in the strong sense* iff all tuples from $J_1 \times ... \times J_n$ exhibit a pluralism of comprehensive doctrines in their respective *PC*-subsocieties.
- *of grade r* iff a proportion $r \in [0, 1]$ of all tuples from $J_1 \times ... \times J_n$ exhibits a pluralism of comprehensive doctrines in their respective *PC*-subsocieties.

Note that the notion of a *PC*-subsociety is *tuple-relative*, because only in a specific tuple can be said which agents accept *PC*. In other tuples, these

agents might not accept *PC*, but perhaps others do. Suppose there is a potential local overlapping consensus in the strong sense. Then it is correct to say: For every tuple of justified belief systems, there is a set of agents accepting the same *PC* but a pluralism of comprehensive doctrines. But it might (!) be incorrect to say: There is a set of agents such that for every tuple of justified belief systems, these agents accept the same *PC* but a pluralism of comprehensive doctrines.

This is an important point to realise when it comes to interpreting these notions of an overlapping consensus. I think the best way to interpret a potential local overlapping consensus is to connect it to some notion of compatibility as gestured at above. The main challenge of this thesis and of Rawls's PL is the worry that pluralism and consensus might sometimes not go well together in justified belief systems. If there is a potential local overlapping consensus, even if it's just in the weak sense, then this worry is somewhat alleviated. In fact, it seems plausible to say that the worry is alleviated more so if there are *many* tuples that exhibit a pluralism in the *PC*-subsociety, i.e. there is a potential local overlapping consensus of a high grade.

However, the question remains which practical consequences can be drawn from the existence of a potential local overlapping consensus. Again, as of now, we can talk only abstractly about these things. Suppose there is a potential local overlapping consensus in the weak sense. Thus, there is a tuple from $J_1 \times ... \times J_n$ such that some subset of agents agree on *PC* while endorsing a pluralism of comprehensive doctrines. What can we do to turn this into a potential *global* overlapping consensus? The most straightforward idea is to isolate the favourable conditions of the subsociety and bring about these conditions in the rest of scoiety as well. (This 'bringing about' should not be problematically *forced*, of course.) The goal of this strategy is that there is a new tuple of justified belief systems with pluralism and consensus *across the board*, not just in the *PC*-subsociety. Again, I will flesh out more details of how this might work in chapter 6. Let's suppose this worked. Then this gives us a potential *global* overlapping consensus in the weak sense. That is a step forward, yay!

Can we also turn a potential local overlapping consensus *in the strong sense* into a potential *global* overlapping consensus in the strong sense? Here

we run into the problem mentioned above that the set of agents accepting *PC* is tuple-relative (because acceptance of *PC* is tuple-relative). In some tuple it might be agents $A_1 \subset A$ whose belief systems exhibit pluralism and consensus, in some other tuple it might be agents $A_2 \subset A$ with $A_1 \cap A_2 = \emptyset$. There is just no conceptual guarantee that there is a fixed set of agents that we can look to in order to find favourable conditions that will ensure that *all* tuples exhibit pluralism and consensus across the board (once we bring about these favourable conditions in the rest of society). As of now, we can say nothing more interesting. That is, as of now, a potential overlapping consensus in the strong sense is worth as much as a potential overlapping consensus in the weak sense, at least when viewed from the practical perspective. Of course, it might very well turn out that in any particular real society, it's typically the same set of agents with consensus and pluralism in all or many tuples. Then we have a reference point that we can look to so that we can bring about conditions that give us a potential global overlapping consensus of high grade.

Note that I have not talked about the notion of an *actual* local overlapping consensus (even though its definition would be straightforward). The reason for this is that I simply don't see its relevance over and above its implying a potential local overlapping consensus.

Let's recap this section. Starting from Rawls's idea of an overlapping consensus as the gold standard for stability in a society, I have given a semi-formal definition that captures this idea: the actual global overlapping consensus (definition 1). This definition presupposes a holistic notion of justification that will be motivated in the next section. I have also discussed two variations of this definition:

1. I proposed the notion of a *potential* global overlapping consensus (definition 2). The modal 'potential' indicates that even though citizens might not actually hold justified beliefs, the belief systems that would be justified for them nonetheless do form an overlapping consensus. That is, if citizens *did* hold justified moral belief systems, there *would* be an (actual global) overlapping consensus. However, for any citizen there might be several different justified belief systems, leading to a host of possible combinations of these belief systems in their society. Thus, we have to differentiate *different senses* of a potential global overlap-

	Local	Global
Potential	Definition 4	Definition 2
Actual	(irrelevant)	Definition 1

ping consensus. If all combinations of justified belief systems exhibit pluralism and consensus, then there is a potential global overlapping consensus *in the strong sense*. If there is at least one such combination, then *in the weak sense*. If there is a proportion of $r \in [0, 1]$ of such combinations, then *of grade r*. These different senses correspond to different conditional probabilities for there to be an actual global overlapping consensus given that citizens hold justified belief systems.

2. I proposed the notion of a potential *local* overlapping consensus (definition 4). Here the idea is that even though there might not be society-wide consensus and pluralism in a given combination of justified belief systems, there might nonetheless be a *subsociety* for which this holds. If there is, then this indicates that in this society consensus on the political conception is in some sense *compatible* with pluralism of comprehensive doctrines. Again, since there can be many combinations of justified belief systems in a society, I have differentiated different senses of this potential local overlapping consensus in direct analogy to the global kind.

Regarding these definitions, I have made explicit how their successful application to a particular society can be interpreted and what practical consequences might be drawn.

2.2 Reflective Equilibrium

One main takeaway of the last section was that citizens need to be morally justified in endorsing the political conception. The goal of this section is to get a first grip on what 'being justified' can mean in this context. Again, I first present Rawls's take on the matter, before developing my own commitments out of discussing his.

2.2.1 Equilibrationism

Many philosophers, including Rawls, believe that the relevant method of justification in ethics and political philosophy is the *method of reflective equilibrium* (MRE). In his *A Theory of Justice*, he discusses MRE explicitly and somewhat extensively as a method of justification (TJ, sections 4 and 9; see also Rawls, 2001). In *Political Liberalism*, in contrast, the method itself is not discussed in detail. However, it is clear that Rawls still endorses the method: It is referenced throughout the book, though sometimes with shorthand phrases like 'upon due reflection' and otherwise loose terminology, e.g. talking of 'acceptability' or 'reasonability' instead of justification. Thus, TJ is a better source for an exposition of Rawls's view on MRE.

As he presents the view in TJ, MRE is a method for justifying moral theories, or moral convictions in general. The general idea goes like this (cf. TJ 18f.): One compares a theory to one's considered judgments about the subject matter. If there is a misfit, either can be revised. By going back and forth between the two levels and adjusting one to the other, one eventually reaches a *state of reflective equilibrium*. In this state, both levels fit and form a coherent whole of mutually supportive considerations. A theory is justified to the extent that it is the result of such a process, or can be rationally reconstructed as such. In fact, as I foreshadowed in the last section, MRE is a holistic approach to justification. The fact that both levels fit means that they support *each other*. Thus, if one is in reflective equilibrium, then not only one's moral theory but the moral belief system as a whole is justified.

Rawls proposes a certain expository device that is supposed to help in the equilibration process: the *initial situation* (TJ 19). This is a hypothetical scenario in which some agents collectively choose principles of justice. This choice problem can be described in various ways and contain various assumptions about the motivation, knowledge, and so forth, of the agents. Given some description, the agents will choose a utilitarian principle, given some other description, they will choose Rawls's JF. Of course, Rawls is particularly concerned with spelling out the latter description, which he calls the *original position*, containing amongst others his famous veil of ignorance (cf. TJ 102ff.). Given this picture, MRE will bring three pieces into equilibrium: the description of the initial situation, the principles that would be chosen in

this situation, and our considered judgments. Using the expository device of the initial situation, we are going back and forth not between judgments and theory directly, but between judgments and the description of the initial situation, which in turn yields principles which in turn fit or do not fit our judgments.

It is somewhat of an open question whether the initial situation is in fact *just* an expository device or whether and how it contributes to justification. For example, one could say that we choose some description of the initial situation over another not only because it leads to a good fit between theory and judgments, but also because it is 'inherently' more plausible than the alternatives. It's not entirely clear what Rawls's stance on this is, but I myself am very skeptical. That is, I don't think that the initial situation contributes to justification. All that matters is the fit between theory and judgments. The reason for my skepticism is the following. The initial situation is clearly a tool from the social contract tradition of moral theorising. Thus, one might ask: Is it fair to build such a tool into the definition of moral justification? I submit that it is not fair . The initial situation should count at most as an *optional* expository device that does not itself contribute to justification. Otherwise, a dialectical opponent can complain: 'Sure, *if* you use the initial situation in the equilibration process, *then* you have the upper hand. But if we compare theory and judgements directly, then my theory is all things considered more plausible.' A related objection complains that, because the initial situation always abstracts somewhat from the actual circumstances, some injustices (e.g. concerning gender and race) become hard or even impossible to address (Pateman, 1988; Mills, 1997). Since I see no independent reason for *requiring* us to use the initial situation during equilibration, I think that the initial situation should at most count as an optional expository device. What really and exclusively matters for justification is that theory and judgements are in reflective equilibrium.

To be sure, if the initial situation is a useful cognitive tool that helps us think about these matters, and perhaps helps us arrive at a coherent belief system, then that helps justification in a sense, but only with respect to the *genesis* of the justified beliefs, i.e. their actually coming about. It does not follow that the initial situation should be part of the *criterion for their justifiedness*. This distinction between genesis and justifiedness occurs promin-

ently in philosophy of science as Reichenbach's so-called *context distinction* between context of discovery and context of justification (Reichenbach, 1938, pp. 6f), but can likely be tracked further back in the history of philosophy and applies equally to epistemology in general (Hoyningen-Huene, 1987). Since I am interested in the justifiedness of the citizen's beliefs, and not their genesis, I will henceforth ignore the initial situation. In particular, it should play no role in defining what it takes for a belief system to be justified. This point is important and will be further explained in section 2.2.3.

Two qualifications to this point. First, I only reject the idea that the initial situation contributes to justification by being a third piece, next to theory and considered judgments, that must be brought into equilibrium during an equilibration process. That is, I reject Rawls's idea that citizens do not *directly* adjust theory and considered judgments to each other, but instead *indirectly* via adjustment of the description of the initial situation. What I don't reject, however, is that an initial-situation-style argument for a particular theory can contribute to the justification of this theory. In the study design in chapter 4, such arguments are represented in the same way as any other argument. Second, even though I generally reject the idea that the initial situation contributes to justification (in the way described above), I am strictly speaking only committed to rejecting that the initial situation contributes to what Rawls calls full justification, as opposed to pro tanto justification or public justification. This comment will become clear only in the next section 2.2.2, where I discuss these notions.

Thus, what we are left with is the characterisation of MRE I gave in the beginning of this section: We adjust theory and considered judgements to each other until the two levels form a coherent whole. Note that this account of justification is neither purely foundationalist nor purely coherentist (cf. Schmidt, 2022). It is not purely foundationalist, because it does not presuppose a certain and unrevisable basis of beliefs (i.e. the foundation) from which the rest can be inferred (as, paradigmatically, Descartes imagined). In the adjustment process, no commitment is sacrosanct, everything can be revised. At the same time, this account of justification is not purely coherentist, because coherence is not *all* there is to justification. In particular, since the adjustment process starts from a given set of considered judgments, there is a tie to these judgments. As a consequence, not any coherent belief system is

justified, but only those that result from such a process (or can be thus reconstructed). This answers the objection to pure coherentism that any coherent belief system, no matter how absurd, counts as justified (cf. Olsson, 2023, §1). As a consequence, this account of justification can best be described as *weakly foundationalist* (Elgin, 2005; Beisbart and Brun, 2024; Schmidt, 2022; Rechnitzer, 2022; terminology by BonJour, 1985; for an opposing view see Tersman, 1993).

Given this characterisation, there are two immediate follow-up questions: First, what are considered judgements? Second, what does it mean to say that theory and considered judgements form a coherent whole?

What are considered judgements?

I should start by saying that the notion of 'considered judgement' plays a subtle double role in Rawls's account of justification. First, considered judgements are the starting point for the method of reflective equilibrium. As such they give an important reference point for equilibration and justification. In particular, this means that a belief system is always only justified *relative to* a set of considered judgements. Second, considered judgements change and evolve during equilibration. They are a part of the agent's justified belief system and as such they are what fits into a coherent whole once a state of reflective equilibrium has been reached. Thus, they are the justified beliefs or commitments of the agent. This double role might seem obvious and uninteresting right now, but it is important to realise early on that these are two different functions that might, in principle, be fulfilled by different entities. In fact, the model of MRE presented in chapter 3 does assign different entities for these two roles.

Of course, the first role of considered judgements as the starting point for MRE requires clarification. Where does this starting point come from? According to Rawls, our considered judgements are judgements "rendered under conditions favorable to the exercise of the sense of justice, and therefore in circumstances where the more common excuses and explanations for making a mistake do not obtain" (TJ 42). By 'sense of justice' Rawls means our "skill in judging things to be just and unjust, and in supporting these judgments by reasons" (TJ 41). Though influential, this rather demanding

characterisation is, of course, not the only possible one. In particular, some authors endorse much more permissive characterisations, as Elgin does with her notion of "initially tenable commitments" (Elgin, 2017, p. 64; also see, e.g., Lewis, 1983, p. x).

I will not here dive into this matter, even though it is of central philosophical importance and considerable conceptual difficulty. The reason for my neutrality is that the research presented in the present thesis is completely independent of this question. All I am committed to is that there is *some* plausible starting point for MRE such that the results of an equilibration process can count as justified. I also assume that this reference point can plausibly be represented by a set of sentences (see section 3.2). Whoever shares this basic assumption might find the results of the present thesis relevant. However, their exact interpretation and, importantly, their practical consequences will depend on a precise account of this starting point. If you need some idea of what might be such a starting point, you can rely on the above Rawlsian characterisation or simply think of this starting point as the moral intuitions of the agent.

What does it mean to say that theory and considered judgments form a coherent whole?

So far I have left open what the criterion for the two levels 'fitting together', or 'forming a coherent whole', or 'mutually supporting each other', is. Of course, there is a host of literature on how to characterise coherence (for some seminal contributions see Ewing, 1934; BonJour, 1985; Shogenji, 1999; Thagard, 2000; Bovens and Hartmann, 2003). However, I will not survey this literature here and instead focus on what will be the basis for the explications in the next chapter.

Two core concepts will help us get a grip on coherence: *derivability* and *systematicity*. The following quote of Rawls brings this out nicely:

> "[W]hat is required is a formulation of a set of principles [i.e. a theory] which, when conjoined to our beliefs and knowledge of the circumstances, would lead us to make these judgments [i.e. the considered judgements] with their supporting reasons were we to apply these principles conscientiously and intelligently.

> [...] These principles can serve as part of the premises of an argument which arrives at the matching judgments. We do not understand our sense of justice until we know in some systematic way covering a wide range of cases what these principles are." (TJ 41)

Rawls thinks that in reflective equilibrium there is an asymmetry between theory and judgements, namely, that they can be connected by arguments in a certain way. In these arguments, the theory (the 'set of principles') appears in the premises, perhaps together with some premises concerning the circumstances, and the judgments appear in the conclusions. In that sense, the judgments are *derivable* from the theory. Rawls also thinks, or it is at least one salient interpretation of this quote, that the principles should in some *systematic* way cover a wide range of cases. This point is important: It is not enough to simply list your considered judgments and call it your theory, as Rawls himself stresses (TJ 41). In that case, the judgments would still be (trivially) derivable, but that theory would not be systematic or systematise the judgements.

I fully agree with Rawls regarding both derivability and systematicity and, as we will see later, I am substantially committed to these ideas. The formal model of reflective equilibrium presented in section 3.2 offers an explication of these notions. For now, we can stick to the following informal characterisation: A belief system forms a coherent whole iff it contains a systematic theory from which the (other) commitments of the agent are derivable. I used the term 'commitments' here to avoid presupposing the Rawlsian double role of considered judgements as both the starting point for MRE and the commitments that evolve during equilibration. In what follows I will often use the term 'initial commitments' to denote the starting point and 'commitments' to denote the (evolving) beliefs of the agent.

This concludes my first informal characterisation of the method of reflective equilibrium. The preliminary result can be summarised as:

> **Equilibrationism** A belief system is justified iff it is or can be reconstructed as the result of an equilibration process in which theory and commitments are adjusted to each other until they form a coherent whole. That is, they form a belief system in

which the commitments can be derived from a systematic theory.

Many philosophers have subscribed to the idea that this or some similar version of equilibrationism gives the correct criterion for the justification of moral beliefs (sometimes even philosophical or scientific beliefs in general), see Daniels (1996); Elgin (1996); Scanlon (2003); DePaul (1993); Lewis (1983); Beauchamp and Childress (2013); Doorn (2010); Mikhail (2011); Swanton (1992); van der Burg and van Willigenburg (1998); Keefe (2000).

Let's recap this section. I presented the Rawlsian characterisation of the method of reflective equilibrium and committed to the basic picture: Theory and commitments are adjusted to each other until they form a coherent whole. In particular, I have committed to his idea that coherence of a belief system means (at least in part) that the theory systematises the commitments. I am not, however, committed to his conception of the starting point of this process, which he calls considered judgements. The model presented in the next chapter is compatible with many different such conceptions.

There still are some open questions that I will address soon. Before we turn to these issues, however, I wish to clear up an important point that might be on the mind of anyone who is somewhat familiar with Rawls's political liberalism: The relation between public reason and justification.

2.2.2 Full justification, public reason and assurance

In section 2.1.1 I said that I am not committed to the idea of public reason. Yet public reason plays an important role in Rawls's ideas about justifying the political conception of justice on which the comprehensive doctrines overlap. In fact, much of the debate on political liberalism has focused on this point (see below in this section). Thus, how can I talk about justification without talking about public reason? Answering this question will not only back up my bold refusal to give public reason a prominent stage in this thesis. It will also help understand what the present project is and is not supposed to be. My answer has two parts. First, I discuss Rawls's distinction between three kinds of justifications of the political conception. Second, I situate my project with respect to the recent discussion about how public reason can solve the so-called assurance problem.

Three kinds of justification

Public reason, according to Rawls, is a part of the political conception of justice, next to the substantive principles of justice he defended in TJ (e.g. PL 224f, but also see his later refined view on public reason in Rawls 1997). It gives citizens (including, importantly, government officials) the resources to reason about questions of basic justice without relying on any particular comprehensive doctrine. In fact, the principles of justice themselves can be justified by public reason. However, this is not the kind of justification I am interested in. Rawls himself disinguishes three kinds of justifications of the political conception (cf. PL 386f):

- *Pro tanto* justification: The political conception is justified using the resources of public reason alone, i.e. referencing only political values. In terms of MRE, public reason is in reflective equilibrium with the principles of justice. (Since principles and public reason make up the political conception, in a sense the conception itself is in RE.) The justification is *pro tanto* because, in principle, the political values may be overriden by non-political ones once the political conception is not considered in isolation, but in a wider view.

- Full justification: A political conception is fully justified by an individual citizen if it is embedded in their comprehensive doctrine. That is, the different parts of the comprehensive doctrine and the political conception are in reflective equilibrium. If all citizens have in this sense fully justified the political conception, then there is an overlapping consensus, or consensus for the right reasons. To use another Rawlsian notion, they are in *full reflective equilibrium* (PL 384n).

- Public justification: A political conception is publicly justified by a political society (as a collective, not individual citizens). This public justification, as Rawls imagines it, is based on there being an overlapping consensus and on the idea of stability for the right reasons (see section 2.1.2) and the principle of legitimacy (PL 388f). The details need not concern us here, the important point is that public justification is both different from and dependent on the full justification of individual citizens in an overlapping consensus.

Note that both *pro tanto* justification and public justification belong to the *public sphere*, because they must not presuppose any particular comprehensive doctrine. Full justification, on the other hand, may do so and thus belongs to the *non-public sphere*. Also note that not only public justification depends on there being a society-wide full justification of, or overlapping consensus on, the political conception. *Pro tanto* justification, too, depends on this to the extent that it is supposed to be appealing to all citizens, because it uses the resources of public reason and public reason is a part of the political conception (at least according to Rawls). If there is no overlapping consensus on the conception, a reference to public reason will not be convincing to all citizens and the *pro tanto* justification will not be public in the proper sense. Thus, society-wide full justification is a precondition for the other two kinds of justification, if these are supposed to appeal to all citizens.

Now, for the purpose of this thesis I am interested in this precondition, in the possibility of an overlapping consensus. That is, I am interested in the possibility of all citizens having fully justified the same political conception even though they endorse a variety of comprehensive doctrines. In particular, the two justifications of the public sphere and the structure and content of public reason are not directly relevant here. Thus, public reason itself will not be modelled in chapter 4. (Nonetheless, it is in a weak sense represented, because the political conception, of which it is a part, does appear as an entity in the model.) In essence, my focus is on the non-public sphere with the corresponding notion of full justification and not the public sphere with the corresponding notions of public reason and public justification. I will return to this issue (non-public sphere vs. public sphere) in section 2.2.3 where I discuss the influence of the public political culture on the formation of an overlapping consensus.

A final remark on this clarification: In the last section I argued that the initial situation should not count as contributing to justification, because there is no independent reason for requiring us to use this device. It now becomes clear that I am only committed to holding this view with respect to individual, non-public, full justification, as I foreshadowed in the last section. So even if you think that there are decisive reasons for requiring us to use the initial situation when giving, e.g., a *public* justification for some political conception of justice, that's still compatible with the assumptions

for the present project which is about full justification.

Public reason and the assurance problem

Let's turn to the recent discussion about public reason, because this discussion has been very prominent and it is closely related to the idea of an overlapping consensus. I should note that the debate does not often differentiate sharply between public justification and justification by public reason (as Rawls does above). For example, the SEP articles on public reason and public justification, respectively, mostly agree both in identifying critical issues as well as the relevant positions regarding these issues (Vallier, 2022; Quong, 2022). In what follows, I will present the discussion about the assurance problem by using the using the term 'public reason' and not 'public justification'.

Many publications regarding public reason start by referencing Weithman's (2010) analysis of the Rawlsian political turn. As mentioned in section 2.1.1, Weithman thinks that Rawls was mainly motivated by the problem of stability in pluralist societies for which he offered the idea of overlapping consensus as a solution. (I agree with Weithman's reconstruction and presented this issue accordingly.)

Additionally, however, Weithman explains that Rawls's account of stability is not exhausted by the idea of an overlapping consensus. Even if there is such a consensus, citizens still face an *assurance problem* (Weithman 2010, §§II.1–3, Weithman 2015, pp. 83f): How can I be sure that I am not the only one who upholds the conception of justice while the others don't? This is, Weithman analyses, a form of the classic prisoner's dilemma from game theory. At this point, public reason comes into play. Rawls proposes to solve the assurance problem by requiring citizens to use the resources of public reason to deliberate about political matters, in particular, about matters concerning constitutional essentials. Since public reason is a part of the shared political conception of justice, citizens thus signal adherence to this conception. As a consequence, it is public knowledge that there is an overlapping consensus and the assurance problem is solved.

This point sparked much discussion. In particular, there is a big debate on whether citizens really have to use shared reasons in public deliberation

about constitutional essentials (as Rawls originally required, though later qualified (1997)) or whether they can also use their own non-public reasons and still assure each other of their allegiance to the shared conception of justice. The former view is called the *consensus view* on public reason, while the latter is called *convergence view*. For example, Hadfield and Macedo (2012) or Wong and Li (2023) can be put, roughly, into the consensus camp, while Gaus (2011), Thrasher and Vallier (2013), Kogelmann and Stich (2016) or Kogelmann (2019) belong to the convergence camp. As you can imagine, I do not wish to enter the debate and, as of now, I see no need to. Both views presuppose that there is an overlapping consensus, they only disagree on what it takes to solve the subsequent assurance problem. In a sense, I am concerned with the precondition for this debate, namely, with what it takes for there to be an overlapping consensus such that there is an assurance problem to begin with. Quong (2011, ch. 6) stresses this point, i.e. that an overlapping consensus is conceptually prior to public reason, and I agree with him in that regard.

Of course, the two camps may disagree about what needs to be in the focus of an overlapping consensus, i.e. what a political conception of justice must include. In particular, the consensus view requires that the political conception includes a full account of public reason such that all citizens can draw from this pool of shared reasons when arguing about constitutional essentials. The convergence view does not necessarily rely on such an assumption. But, as I have stressed multiple times already, I am not concerned with what is or isn't in the content of a political conception of justice, because I adopt a purely structural perspective. Thus, philosophers from both camps can find the present investigation to be of interest to them.

The same holds for the closely related discussion about public reason vs. public deliberation (Vallier, 2015; Boettcher, 2020; Kugelberg, 2021). Very roughly, the question here is more generally whether citizens have to use any kind of public reason or whether it suffices if government officials do so. Again, it seems that at least some of the participants to this debate presuppose that there is an overlapping consensus, thus, my research will be of interest to them. At least, my research does not conflict with either position on the matter.

The bottom line is that an account of societal stability might require more

than an overlapping consensus. Perhaps it also requires a solution to the assurance problem. The debate I have just outlined presupposes that an appropriate account of public reason can (and perhaps must) contribute to a solution. I myself am not entirely sure how pressing the assurance problem really is. Part of my skepticism is grounded in my agreement with Klosko (2015) that in real societies stability is always in part 'imposed' (see section 2.1.2). That is, to some extent I can expect my fellow citizens to adhere to the constitution, simply because the state will punish transgressions of it. (Of course, the participants to the above debate do not deny this, but they are interested in purely inherent stability, especially when concerned with traditional ideal theory.) Nonetheless, I do acknowledge that mutual assurance in the form of using public reason might be very helpful for social cohesion and stability. But my research on the realisability of an overlapping consensus is independent of this debate, because, first, the existence of an overlapping consensus is prior to the assurance problem and, second, I adopt a purely structural perspective, thus, the results will be of interest to all, no matter what they think about the locus and the content of public reason.

Let's recap. In this section, I have argued that I can investigate the realisability of an overlapping consensus without giving public reason a prominent stage or modelling it in the simulation studies. First, I have cleared up that by 'justification' I mean what Rawls calls 'full justification': Citizens fully justify a political conception by bringing it into reflective equilibrium with their comprehensive doctrine. If all citizens do this, then there is an overlapping consensus. This overlapping consensus is a precondition for the proper use of public reason and for giving a public justification (in the above Rawlsian sense) of the political conception. Second, I have stressed that it is also a precondition for the so-called assurance problem. Only if there is an overlapping consensus can we ask how to assure each other of this consensus. Even if a solution to the assurance problem is important for societal stability and public reason is important to such a solution, the purely structural perspective allows me to remain agnostic about these matters, since the results are compatible with and relevant for all sides of the debate.

2.2.3 Reconstructionism

There is another important Rawlsian concept that will play no role in what is to come, though it may play a role in follow-up studies: the *public political culture* of a society. For one thing, I wish to argue that it is legitimate for me to ignore this concept for the time being. For another, this discussion will uncover an important commitment about equilibrationism that I share with Rawls, namely *reconstructionism*.

The public political culture of a society comprises "the political institutions of a constitutional regime and the public traditions of their interpretation (including those of the judiciary), as well as historic texts and documents that are common knowledge" (PL 13f). One might think that growing up in and living exposed to a public political culture will have an influence on the likelihood of accepting the political conception that is realised in this culture. In fact, when explaining how an overlapping consensus might come about (PL §§6-7), Rawls speculates that living in a public political culture might be a crucial driving force behind the formation of an overlapping consensus:

> "This suggests that many if not most citizens come to affirm the principles of justice incorporated into their constitution and political practice without seeing any particular connection, one way or the other, between those principles and their other views. It is possible for citizens first to appreciate the good those principles accomplish both for themselves and those they care for, as well as for society at large, and then to affirm them on this basis. Should an incompatibility later be recognized between the principles of justice and their wider doctrines, then they might very well adjust or revise these doctrines rather than reject those principles." (PL 160)

It seems that one needs to take the public political culture in a society into account when investigating the possibility of an overlapping consensus. In fact, since the use of public reason is part of the public political culture, this would mean that public reason is, contrary to what I said in the last section, very important for the realisability of an overlapping consensus.

However, one must not conflate two fundamentally different aspects of belief: *genesis* (whether and how beliefs come into existence) and *justifiedness*

(whether and how beliefs are justified). Both aspects are important, because we would like the beliefs in an overlapping consensus to be both existent and justified. In the above quote, Rawls describes the role of public political culture in the genesis of an overlapping consensus. This thesis, however, is concerned with justification and not genesis.

Consider the following two specifications of equilibrationism:

> **Actualism** Beliefs are justified iff they are the result of an equilibration process. (MRE describes the actual process of generating justified beliefs.)
>
> **Reconstructionism** Beliefs are justified iff they could have been the result of an equilibration process. (MRE is a test for whether beliefs are justified, no matter how they were generated.)

An epistemologist's acceptance of either specification will depend on what they take MRE to be. I think that, as a general rule of thumb, if MRE is taken as a rather wide umbrella term for all kinds of processes that may lead to a state of reflective equilibrium, then Actualism may be a viable position. But when MRE is taken to be a rather precisely specified algorithm, as I will do in chapter 3, then the claim should be weaker and some version of Reconstructionism. Rawls himself, after describing MRE as the step-wise mutual adjustment of theory and judgements, subscribes to Reconstructionism:

> "I shall not, of course, actually work through this process. Still, we may think of the interpretation of the original position that I shall present as the result of such a hypothetical course of reflection." (TJ 18)

Other philosophers have also subscribed to some version of Reconstructionism (Goodman, 1955; Elgin, 2017; Baumberger and Brun, 2021). I, too, interpret MRE as a test for the justifiedness of beliefs, not a method that citizens must actually apply in their belief dynamics. Thus, since the public political culture plays its role in the generation of citizens' beliefs, but I am concerned with using MRE as a test for the justifiedness of the beliefs, I will not model a public political culture when simulating equilibration processes.

Nevertheless, one might say that sharing a political culture is relevant not only for the genesis but also for the justifiedness of the citizens' beliefs.

For example, sharing a political culture might lead to the citizens' initial commitments being similar in a certain way. This would be relevant for the justification of belief systems, since the initial commitments are an important reference point for equilibration processes, both actual and reconstructed. This can be easily modelled, though the study design presented in chapter 4 does not have this feature. In chapter 6, I present a suggestion for capturing this aspect in future studies.

My commitment to Reconstructionism also clears up an issue that might have been on your mind for a while now: Is this thesis concerned with propositional or doxastic justification? Roughly, propositional justification only requires that the agent has good reasons or evidence for believing a proposition. Doxastic justification additionally requires that the agent believes the proposition on the basis of their reasons or evidence. (The distinction goes back to Engel (1992), who called the former 'personal justification', and has since been extensively discussed by epistemologists. For a survey on the epistemic basing relation involved in doxastic justification, see Korcz (2021).) The following example illustrates the distinction:

> "Imagine two jurors, Miss Knowit and Miss Not, deliberating about the case of Mr. Mansour. Both jurors have paid close attention throughout the trial. As a result, both have good reason to believe that Mansour is guilty. Each juror goes on to form the belief that Mansour is guilty, which he in fact is. Miss Knowit believes he's guilty because of the evidence presented during the trial. Miss Not believes he's guilty because he looks suspicious." (Turri, 2010, p. 312)

For both Miss Knowit and Miss Not, it is propositionally justified that Mansour is guilty, but only Miss Knowit's belief is doxastically justified, because Miss Not did not form her belief on the basis of her evidence.

It is not entirely straightforward how to apply the distinction between propositional and doxastic justification to the present equilibrationist account of justification, because there is no explicit mention of reasons or evidence. But according to Reconstructionism, in contrast to Actualism, justification is not sensitive to how the beliefs were generated. As a consequence, it seems that a reconstructionist interpretation of equilibrationism

constitutes an account of *propositional justification*, at least as long as there are no additional requirements pertaining to how the agent's beliefs were formed. Since I will not introduce such additional requirements, the present thesis is concerned with propositional, not doxastic justification.

Note, however, that the explication of MRE offered in the next chapter 3 operates with sentences, not propositions. Thus, strictly speaking, the explication of justification will be one of sentential justification. But this terminology is non-standard and I will instead use the more familiar term 'propositional justification'. Also note that the difference between these kinds of justification is usually thought to be associated with who or what counts as justified. Is it a proposition, a person or a belief? And this, in turn, can be reflected by how statements about justification are phrased: Miss Knowit's belief that p is (doxastically) justified while the proposition that p is (only propositionally) justified for Miss Not. Thus far, I have not been consistent with who or what is justified: Sometimes I have talked of agents being justified in holding a belief system, sometimes of justified belief systems, sometimes of a political conception that is justified for an agent. I will continue to be sloppy about this, mainly for ease of exposition, but also because I don't think that the differences in natural language expression semantically track this philosophical distinction. In case of doubt, I should always be understood as saying that some set of sentences is justified for an agent.

You might wonder whether propositional justification is really what we are after when it comes to an overlapping consensus account of stability. It might seem that citizens' affirmation of a political conception is *particularly stable* if they do not only have good (and moral) reasons to accept the conception, but also accept the conception on the basis of these reasons. Does the present account of justification miss the mark? There are at least two possible responses to this objection. First, one can argue that propositional justification is sufficient for a stable affirmation of the political conception. We can proceed with the present conceptual toolkit. The second response, which I am inclined to give, is to admit that doxastic justification brings more stability to the table than mere propositional justification. But it should also be stressed that propositional justification seems to be a necessary condition for doxastic justification. On the standard account of doxastic justification

(mentioned above), an agent's belief in p is doxastically justified iff p is propositionally justified for the agent and the agent believes p on the basis of the reasons that constitute p's propositional justification. Thus, no doxastic justification without propositional justification. (On non-standard accounts like Turri's (2010), things are less clear.) If this is correct, then studying conditions for the weaker form of overlapping consensus (involving propositional justification) also informs us about conditions for the stronger form of overlapping consensus (involving doxastic justification). (In effect, this adds a further distinction to the ones given in section 2.1.3.) If there are conditions such that not even the weak form is possible, then the strong form is likewise off the table. However, if the weak form *is* possible, then this is not yet a guarantee that the strong form is possible as well. In order to ascertain this, we would have to give a precise equilibrationist account of doxastic justification. This account will likely involve requirements concerning *how* citizens form their beliefs. We would then have to investigate how the actual opinion dynamics of the citizens fare with respect to these requirements.

In essence, these considerations once again highlight the limited scope of the present research. It is not about finding conditions that must hold for the actual belief dynamics of the citizens such that an overlapping consensus comes about. Instead, the question is: In how far does the requirement that citizens in an overlapping consensus are propositionally justified limit the possibility of an overlapping consensus?

2.2.4 Epistemic consequentialism and bounded rationality

The term 'reflective equilibrium' can mean two things. It can mean the *method* of reflective equilibrium or the *state* of reflective equilibrium. The latter may be a matter of degree, call these 'degrees of being in the state of reflective equilibrium', short *degrees of equilibrium*. That being in the state of reflective equilibrium is a gradual notion seems plausible not least for the fact that coherence, which is an important feature of being in equilibrium, admits for degrees.

In this section, I wish to adress an important question, namely, whether the method or the state of reflective equilibrium gives the relevant criterion for evaluating one's epistemic state as justified. To be sure, Rawls imagines

that the method of reflective equilibrium always leads to a state of reflective equilibrium, at least for the context that he intends to use it (TJ 18). Nonetheless, it is possible to specify method and state of reflective equilibrium independently from each other. In that case, the result of the method might not be a state of reflective equilibrium. In fact, the formal model of MRE presented in chapter 3 does allow for such cases (Beisbart et al., 2021). This raises interesting questions about the relation between method and state of reflective equilibrium. Both seem to give independent verdicts on what epistemic state should be adopted. Which has more authority? There are at least two possible answers to this question:

> **Epistemic Proceduralism** There is a specification of the method of reflective equilibrium such that it has ultimate authority. Whatever belief system results from its application is justified, no matter the system's degree of equilibrium or whether there are systems with a higher degree of equilibrium.
>
> **Epistemic Consequentialism** The degree of equilibrium of a belief system has ultimate authority by giving an axiology for epistemic states. The degree of equilibrium of a belief system is the feature that is deemed epistemically valuable. The method of reflective equilibrium is simply a means to an end, namely increasing epistemic value.

Some proponents of equilibrationism subscribe to some consequentialist claim (e.g. Goodman, 1955, p. 64; and perhaps Rawls, cf. TJ 19). Others seem to explicitly endorse more of a proceduralist stance (e.g. Scanlon, 2014, p. 79). I am strongly leaning towards Epistemic Consequentialism, though I will not argue for it here. Instead, Epistemic Consequentialism will be a general presupposition of this thesis. For further discussion of this issue, see (Baumberger and Brun, 2021, sec. 2.5).

Note that Epistemic Consequentialism as it is stated here is an equilibrationist version of the more general claim that epistemic normativity (in particular, epistemic rationality and justification) is to be understood in terms of epistemic value (e.g. truth). (For an excellent survey of this general idea, see Ahlstrom-Vij and Dunn (2018), for an impressive and explicit application in formal epistemology, see Pettigrew (2016).) Also note that this distinction

between proceduralism and consequentialism is only concerned with the relation between method and state of reflective equilibrium. In particular, I do not want to commit to the claim that the degree of equilibrium is of *intrinsic* epistemic value. It might be that it is (only) instrumentally valuable for a more fundamental, intrinsic value like understanding (Elgin, 1996; Carter and Gordon, 2014, pp. 7f) or truth (as BonJour, 1985, ch. 8, argues for the value of coherence).

What does this mean for justification? The simplest answer is to say that agents are justified iff their belief system has a maximal degree of equilibrium. However, this is an implausibly high standard. Anyone who has ever reflected on moral questions, philosophically or not, knows how difficult it can be to achieve a somewhat orderly and consistent view on morality. To say that anything short of a maximal degree of equilibrium lacks justification is to demand the impossible. In particular, when connecting this demand with the justifiedness requirement of an overlapping consensus, then this gold standard of societal stability is practically useless.

Thus, the question is: How much degree of equilibrium is enough for justification? How much can we ask of real, epistemically non-ideal agents? A plausible answer can be found, I think, if we turn again to the *method* of reflective equilibrium. However, not be reverting to the (in my opinion implausible) proceduralist claim that there is one single method that once and for all gives the relevant criterion for justification. Instead, the consequentialist picture is that there is a host of possible equilibration methods. Some of these are more effective for increasing the degree of equilibrium of one's belief system, some are less effective. And some of them are feasible, some of them are not. The challenge is to find methods that have the best or at least an acceptable balance of feasibility and effectiveness. Suppose we have found such a method. It then seems plausible to say that an agent's belief system is justified iff it could have been the result of this feasible and effective method. To demand more is to ask too much, because we would require that the agent has a belief system that they could only have gotten by using a method that is not feasible for them. To demand less is to ask too little, because they could have used a method that is feasible for them such that their belief system has a higher degree of equilibrium.

In essence, I am embracing a bounded rationality perspective:

> **Bounded Rationality** Epistemic agents are non-ideal. They have limited cognitive resources, etc. As a consequence, an agent's justification only requires that their belief system is (or could have been) the result of applying a feasible and effective method for increasing epistemic value.

The term 'bounded rationality' was coined by Simon (1957, p. 198) in the context of economics, but has since received attention in various fields, including epistemology (e.g. Gigerenzer and Sturm, 2012; Morton, 2017). For a recent broad defense of bounded rationality approaches in epistemology, see Thorstad (2023). Note that Bounded Rationality as it is stated here is geared towards Epistemic Consequentialism above, it is not supposed to capture the general idea behind all bounded rationality approaches.

Of course, which method has an acceptable balance of feasibility and effectiveness will heavily depend on the agent: their cognitive resources, their knowledge of such methods, and perhaps also their other non-epistemic goals. In section 3.3, I present a method of reflective equilibrium that is relatively feasible and relatively effective, or so I argue.

For now, let me stress once more that a bounded rationality perspective is indispensible when investigating conditions for an overlapping consensus that can be an interesting an account of societal stability. If we compare different levels of idealisation with each other, then the more idealised the epistemic agent, the less likely it is that real-world citizens actually satisfy the requirements for justification. But for matters of stability we are interested not in hypothetical justification of ideal agents, but actual justification of real agents. It does not help stability to say: If all citizens were epistemically ideal, then they would form an overlapping consensus given such-and-such conditions. We want to say: Real non-ideal citizens can form an overlapping consensus given such-and-such conditions.

In this section, I have formulated and embraced two epistemological commitments: Epistemic Consequentialism and Bounded Rationality. Epistemic Consequentialism, i.e. the claim that MRE is just a means to the end of increasing the degree of equilibrium of one's belief system, is a presupposition that I will not defend in this thesis. Bounded Rationality, i.e. the claim that agents are only required to use (or could have used) a feasible and effective method for increasing epistemic value, is a commitment that arises

from the goal to find conditions such that real-world, non-ideal citizens can form an overlapping consensus.

2.2.5 Dialectical situations and wide reflective equilibrium

Let's turn to a final conceptual question about MRE: the question which theories and arguments to consider during equilibration. Obviously, the outcome of any feasible and effective equilibration method will heavily depend on this. For example, if utilitarianism is considered as a moral theory during equilibration, then there is a chance that this theory will be chosen during equilibration. If utilitarianism is not even considered, then there is no such chance. In what follows, I discuss Rawls's ideas about this question and make an alternative proposal. This will enable us to make the general research question more precise and develop some testable hypotheses for the simulation study presented later.

A good way to bring out the problem is to imagine that we are using the classical equilibration method of a step-wise adjustment of commitments and theory to each other (see section 2.2.1). If this is the goal, then how does one find a theory in the first place? If the goal of MRE is a good fit between theory and considered judgments, one should not start with a theory that fits very badly right from the start. At the same time, one should not only consider theories that are seemingly close to one's initial commitments. After all, reflection might bring out that a different theory results in a better overall fit considering all relevant philosophical arguments. So which ones of the many logically possible theories and logically possible arguments for them should one consider? It is, of course, impossible to go through all of them, at least for non-ideal agents. But what we might do, Rawls claims, is to "study the conceptions of justice known to us through the tradition of moral philosophy and any further ones that occur to us, and then to consider these" (PL 43). If, after considering these theories, we reach a state of reflective equilibrium, we are in what Rawls calls *wide* reflective equilibrium. This contrasts with *narrow* reflective equilibrium where we only consider theories that seem to fit our initial commitments or a theory that we perhaps already believe, resulting in a mere "smoothing out of certain irregularities" (TJ 43).

I should note that, in addition to considering a wide array of views

and arguments about the subject matter (here: morality), proponents of MRE usually require that we must also consider how the theories under consideration fit with our so-called background theories (Daniels, 1996, ch. 2). Background theories are not about morality but about other, related subjects such as meta-ethics, epistemology, psychology, sociology, etc. The requirement to consider connections to such theories is obviously plausible. And even though Rawls did not explicitly state this as a requirement for MRE, he does respect it in his theorising (Freeman, 2007, p. 40). However, I will neglect this complication in the simulation studies presented later. Not only is it currently unclear to me how to cash out this 'fit' with background theories in the model presented in the next chapter. It would also further add to the already high demand for computational power. As a consequence, I will not model this aspect of wide RE and hope that the general results prove to be robust once this additional requirement is incorporated.

Let's move back to the question regarding which views and arguments in the subject matter of morality (not including background theories) we need to consider during equilibration. I agree with Rawls that narrow RE is not enough and also that it is too much to ask that one consider all logically possible theories. We need some middle ground, some criterion for theories and arguments such that it is plausible to say: An agent needs to consider these during equilibration in order for them to be justified. Let's call these views and arguments the *dialectical situation* of the agent:

> **Dialectical Situation** The dialectical situation of an agent is the totality of views and arguments that the agent has to consider during equilibration such that the outcome can count as justified.

(This characterisation is not meant to be a definition of the term 'dialectical situation'. If it were, then the definition of justification given in section 2.3 would be circular.)

I am not sure, however, whether it is not too much to ask that one consider all theories put forward in moral philosophy, as Rawls requires. After all, not everyone is a philosopher, let alone a moral philosopher. This is a very demanding standard.

My alternative proposal is to focus on *public debate* instead of moral philosophy. Public debate includes contributions in legacy media (TV, print,

radio), parliamentary debate, social media, etc., and plays a central role in democracies (Bächtinger et al., 2018; Lambek, 2024). I suggest that the dialectical situations of the agents is comprised at least of all theories and arguments that are publicly debated in their society. The idea is that citizens should not be able to ignore views and arguments with which they are confronted on a regular basis and which receive considerable public attention. Of course, we are confronted with all kinds of views and arguments, not just the ones that are publicly debated. Who hasn't met a person who couldn't stop babbling about their obscure conspiracy theories? Who hasn't once taken a wrong turn when surfing the internet and was overwhelmed with views and arguments that are weird or outrageous or both? Considering *all* these views in detail would be too much to ask.

But requiring that one consider the theories and arguments that we are confronted with because they are publicly debated seems plausible to me for two reasons. First, if a view is publicly debated, then we are *regularly* confronted with it. If we are regularly confronted with a theory, it seems we cannot as easily dismiss it and just go about our epistemic business as usual. Instead, we have to consider it and reflect on how this view and its arguments fit with our other beliefs. Second, if a theory or argument is publicly debated, then this is an indicator that it should be taken seriously. This point is, perhaps, even more important. Suppose I am regularly confronted with the seemingly weird view of that one annoying friend. This regular confrontation alone might not require me to take the view seriously, if I am at the same time aware that nobody else is even considering it. If, however, I find that the view is indeed publicly debated, meaning that a significant amount of people are spending considerable time to discuss it, then it seems I should take it seriously and consider how it fits with my other views. In effect, my proposal requires a basic form of epistemic trust in public debate: the trust that the publicly debated views are at least to be taken seriously.

To my knowledge, the question of what belongs to one's dialectical situation is underresearched despite its obvious importance: For example, there is ample literature on *peer disagreement*, i.e. on if and how one should change one's belief in the face of disagreement with an equally capable epistemic agent (Christensen, 2007; Kelly, 2010). This debate takes for granted that one should consider the peer's view, perhaps plausibly so, but leaves open when

to consider the views of non-peers or agents with unclear peer status. The debate on *moral disagreement* is not so much concerned with which opposing moral views one needs to consider such that the own can count as justified. Instead, it is mostly about the meta-ethical consequences that can be drawn from the existence of moral disagreements, in particular with whether the fact of moral disagreement itself leads to some form of moral skepticism (e.g. Tersman, 2006; Enoch, 2009). The epistemology of *testimony* asks, very roughly, how deference to judgments of experts (and others) works, epistemically speaking (see Coady, 1992; Shieber, 2015). But the question of dialectical situations is not about deference, it is about mere consideration. The recent discussion about *zetetic norms* (i.e. norms of inquiry) is, amongst others, concerned with norms for evidence-gathering (Flores and Woodard, 2023). But this discussion revolves around whether there are epistemic norms and duties at all (evidence-gathering being an example), how the zetetic relates to the epistemic, etc (Friedman, 2020; Thorstad, 2022). It is not so much concerned with a characterisation of how much evidence-gathering is enough for justification, or how much considering-views-of-others is enough for justification. In essence, I think that 'considering the views and arguments of others' is a rather weak notion (when compared, e.g., to deference) but is nonetheless relevant for justification. Ideal agents, of course, consider all logically possible views and arguments, but there are no such agents. An epistemology for real agents will have to solve this messy question of what is and isn't in such agents' dialectical situations. The philosophical discussions I just scanned will no doubt be relevant for this investigation, but they do not exhaust it.

If my own considerations about dialectical situations and public debate are correct, then we have found a minimal standard for the citizens' dialectical situation that is more plausible than Rawls's idea. Note, however, that there might be a significant overlap of my proposal and Rawls's proposal, since some theories from moral philosophy may be publicly debated as well. But as it stands, my suggestion is that citizens in a society have to consider at least the publicly debated theories and arguments. As I have stressed, it is an open question in how far agents have to consider more than these views and arguments. There certainly are plausible cases of theories and arguments that we have to consider even though they are not publicly

discussed. It depends on the individual and its circumstances what these additional components of the dialectical situation are. Nonetheless, the different dialectical situations of the citizens will have a *common core*, namely the publicly debated views and arguments. It is this common core that I am interested in for the purpose of the present thesis. I am interested in how the common core of the dialectical situations of the citizens influences the possibility of an overlapping consensus.

For this purpose I will make the idealising assumption that the publicly debated views and arguments (the common core) are the *only* ones that citizens need to consider in order for them to be justified by MRE. That is, I make the idealising assumption that all citizens share the same dialectical situation. We may then hope that the results of the study are robust when the respective de-idealisation is made. In fact, the assumption that citizens share the same dialectical situation is the only relevant modelling assumption about dialectical situations that I make. Perhaps Rawls's proposal is even compatible with this assumption. In any case, if you are not convinced by my proposal and instead have a different view on the matter, but agree that it is a reasonable idealisation to suppose that citizens share the same dialectical situation, then the study results in chapter 5 will be of interest to you, though their interpretation will differ from the one I offer in section 6.2.

Note that public debate, on the view presented here, is to a significant extent independent of the use of public reason. In particular, it assumes that the non-public comprehensive doctrines of citizens play an important role in public debate. This does not preclude us from requiring that public reason should be used in certain circumstances, e.g. when discussing constitutional essentials or when certain people or institutions like the president or the judiciary contribute to public debate. But a conception of public reason that requires *everyone* to *always* refrain from citing non-public doctrines and corresponding arguments in public debate is incompatible with the present approach.

Let's recap. In order to specify MRE, we need some idea of what theories and arguments agents need to consider in order for the outcome of an equilibration process to count as justified. These theories and arguments make up the *dialectical situation* of the agent. Rawls supposes that this dialectical situation consists of the theories put forward in moral philosophy (and any

further one that occur to us). I find this standard too demanding. Instead, I propose that for any citizen the dialectical situation consists *at least* of the views and arguments that are publicly debated in their society. I argued that citizens have to take these into consideration, because they are regularly confronted with them and a form of epistemic trust in public debate requires citizens to take them seriously. The dialectical situation probably consists of more, but the publicly debated views form a common core for all citizens. It is the influence of this common core on the possibility of an overlapping consensus that the present thesis is supposed to investigate. For this purpose, I idealise by assuming that the common core is all there is to an agent's dialectical situation, i.e. all agent share the same dialectical situation. Any who think that this idealisation is reasonable will find the results of the present thesis to be of interest, even if they do not agree with my considerations about dialectical situations.

2.2.6 Public debate and overlapping consensus

In this section, I develop a more precise research question that is focused on the influence of the citizens' dialectical situations on the possibility of an overlapping consensus. I formulate testable research hypotheses that can guide the study design. I argue that testing these hypotheses is of great interest for democrats who are in favor of both an overlapping consensus account of stability and a liberal public debate.

Before we start I want to stress once more that the question of the realisability of an overlapping consensus is not well researched. In PL, Rawls himself focuses on reformulating his own theory of justice from TJ (i.e. justice as fairness) as a *political* conception of justice. In particular, he tries to make it freestanding. However, he delimits his ambitions explicitly:

> "The other point of a reasonable overlapping consensus is that PL makes no attempt to prove, or to show, that such a consensus would eventually form around a reasonable political conception of justice. The most it does is to present a freestanding liberal political conception that does not oppose comprehensive doctrines on their own ground and does not preclude the possibility of an overlapping consensus for the right reasons." (PL xlv f.)

Rawls only shows that JF can be freestanding, i.e. it is not on conceptual grounds impossible for different doctrines to overlap on it as a shared module. Importantly, he does not intend to argue in detail that an overlapping consensus is guaranteed or even likely to develop.

Despite these modest statements, Rawls does consider the objection that an overlapping consensus is utopian (PL, Lecture IV, §§6–7) and answers it by outlining one way in which an overlapping consensus might come about. His remarks, however, are rather brief and speculative, as he admits himself. The basic idea is that a liberal democratic constitution is implemented in a step-wise manner, starting from mere electoral procedures and progressing to basic liberties and further elements. Much of his argument seems to rely on the assumption that during this process citizens will 'just see' that living in a liberal democracy is good and thus accept the constitution, revising their comprehensive doctrines if necessary. This is obviously a strong assumption. Nonetheless, there might be some truth to it. In fact, in section 6.2 I make a like-minded, though weaker suggestion when discussing the importance of political participation and civic education.

Despite Rawls's comments on these matters, rigorous and extensive research on the realisability of an overlapping consensus is still missing. As far as I know, political liberalists are just not so much concerned with this point. Instead, they usually theorise about societies in which an overlapping consensus is already realised. An example of this is the debate about public reason that I have discussed in section 2.2.2. But again, the question about the realisability of an overlapping consensus is important. If it turned out that, for whatever reasons, an overlapping consensus is very unrealistic or even impossible, then the Rawlsian solution to the problem of pluralism fails. If, on the contrary, there are realistic conditions that make an overlapping consensus likely, then these might contribute to a guideline for stabilising pluralist societies.

Of course, in computational epistemology (the discipline in which the present thesis is situated) there is a host of research on the components of an overlapping consensus, i.e. on consensus and pluralism. I cannot give a comprehensive review here, but the following are examples I find particularly relevant. Regarding consensus, Freivogel (2023a) as well as Baumgaertner and Lassiter (2023) study in how far reflective equilibrium promotes *con-*

vergence between agents, i.e. in how far it narrows the room for justified disagreement. Regarding pluralism, I wish to highlight the works by Hegselmann and Krause (2002), Singer et al. (2019), and Dorst (2023) on rational *polarisation* which is arguably a form of pluralism or at least something in between pluralism and consensus. (Also the general epistemological debates on peer disagreement (see section 2.2.5) and epistemic permissivism (see section 3.3) are likewise relevant, of course.) In fact, given a fixed subject matter, consensus and pluralism are two sides of the same coin: If there is pluralism, then there is no consensus, and *vice versa*. However, even though there is much work on pluralism and consensus, the present thesis is about a particular *combination* of pluralism and consensus. It is about conditions under which justified agents agree on a political conception of justice whilst disagreeing about morality in general. Thus, the results about pluralism and consensus *simpliciter* do not directly inform us about the combination of it. In a sense, my work can be seen as investigating the rational synthesis of consensus and pluralism, though with a focus on political conceptions and comprehensive doctrines, respectively.

In the last section I said that I am interested in how the common core of the citizens' dialectical situations influence the possibility of an overlapping consensus. But we need to make this more precise: Which feature of this common core am I interested in? The common core will contain the publicly debated comprehensive doctrines as well as the publicly debated political conception(s) of justice. There can be different kinds of inferential relations between each doctrine and each political conception: the doctrine supports the conception, the doctrine is neutral about the conception, or the doctrine is incompatible with the conception (cf. section 2.1.1). The present research is supposed to contribute to uncovering how these inferential connections influence the possibility of an overlapping consensus:

> **Research Question** Which kinds of inferential connections between the publicly debated comprehensive doctrines and a (publicly debated) political conception of justice make a potential overlapping consensus on this conception possible?

Note that this question is only concerned with *potential* overlapping consensus. As I have already mentioned in section 2.1.3, the reason for this is

that I will simulate artificial societies in which I don't have a use for the notion of an *actually* held belief system.

Why is this question in particular, i.e. with its focus on dialectical situations formed by public debate and the inferential connections therein, interesting and relevant? Public debate is a central part of a functioning democracy, for several reasons. First, it fosters citizens' engagement with political issues, encourages them to critically assess policy proposals and enables them to make informed decisions at the polls (Müller and Campell, 2023). Second, it presses governments to make their decisions transparent and holds them accountable for it. Ideally, policymakers use the feedback from public debate to improve their policy (Habermas, 1996, p. 355). Third, public debate is a driver for the overall evolution of concepts, ideas, norms and values (Estlund and Landemore, 2018; Betz, 2014). This way, it can help democratic societies to make long-term progress.

Given that public debate is so important for democracy, we should ask ourselves how to conduct it in the best possible way. One criterion for evaluating public debate is to what extent it helps stabilise a society by fostering an overlapping consensus on the political conception of justice that is implemented in that society (or, alternatively, by bringing out that a different political conception than the one implemented is better at fostering an overlapping consensus on it). If we find features of public debate that are threatening societal stability, e.g. by consolidating instead of bridging divides regarding the political conception, then this gives us a *pro tanto* reason to try to change these features and instead bring about conditions such that public debate fosters an overlapping consensus.

One salient feature of public debate that might have a strong influence on stability is the worldviews or comprehensive doctrines that are invoked or discussed. For example, suppose that the political conception of justice implemented in a society is a liberal and democratic one. But all comprehensive doctrines that are represented in public debate are incompatible with a liberal democratic constitution. They might be doctrines of racial supremacism, religious fundamentalism, etc. Will public debate, under these conditions, foster an overlapping consensus on the liberal and democratic political conception? Probably not. Suppose, on the other hand, that all comprehensive doctrines that are represented in public debate are strongly

supportive of the liberal democratic constitution. It is then much more likely that public debate will foster an overlapping consensus on the liberal democratic political conception. These pro-democratic doctrines will become an important part of all citizens' dialectical situations, making it more likely that they accept one of them and, in turn, accept the political conception supported by them.

The inferential relations between doctrines and and a given conception are a relatively simple and universal feature of public debate. No matter what the doctrines and conceptions of some society are, you can always count the inferential connections between the publicly debated comprehensive doctrines and a given political conception: How many doctrines are supportive of the conception? How many are neutral? How many incompatible? Yet despite this relative simplicity, these connections can be expected to heavily influence the overlapping consensus on that conception. (By 'simplicity' I here mean relative conceptual simplicity. It is, of course, quite difficult to empirically determine these inferential connections in a given society.)

If we find that certain connections (e.g. incompatibility, perhaps neutrality) or combinations thereof make on overlapping consensus more difficult, then we (as citizens in that society) have to discuss how to deal with that. Do we want to exclude such doctrines from public debate? What place, if any, should they have? These are difficult questions. Ideally, a public debate is an open space in which no views and arguments are prohibited and everyone has an equal right to be heard (Habermas, 1990). Limiting this right in whatever form is, to some extent, illiberal and it is unclear how to trade this off against the stability given by an overlapping consensus. The present research is supposed to highlight the cost of allowing views with certain inferential connections to the political conception to be part of public debate.

I want to formulate a testable research hypothesis as a potential answer to the research question stated above. Or rather, I want to formulate several such hypotheses, one for each kind of overlapping consensus I distinguished in section 2.1.3. The purpose of these hypotheses is to have a goal (i.e. to test the hypotheses) that can guide the design of the simulation study (chapter 4) and the analysis of the results (chapter 5). The hypotheses should have some initial plausibility. Or rather, we should have some idea about why

they could be true. More importantly, they should be relevant. That is, it should be important whether they are true or false.

My idea for these hypotheses is as follows: Of the three possible inferential connections that any comprehensive doctrine can have to a given political conception (support, incompatibility, neutrality) there is only one for which it is initially plausible to say that it will foster an overlapping consensus on the conception: the support connection. As described above, if many of the publicly debated doctrines support a political conception, then we can expect this to have a positive influence on the formation of an overlapping consensus, because it will make it more likely that citizens accept such a supportive doctrine. For the incompatibility connection, the reverse is plausible. For neutral doctrines, it is unclear. In particular, there is no reason to think that it will do any good for an overlapping consensus on the conception. Thus, the hypothesis is, roughly, that support connections are necessary for on overlapping consensus.

Let me make this idea more precise in two steps:

1. Is it really necessary that *all* publicly debated comprehensive doctrines support a political conception such that an overlapping consensus on that conception is possible? This seems like a very strong (and implausible) statement. Suppose there are ten publicly debated comprehensive doctrines and nine of them support the conception while one of them is incompatible with it. It seems a bit much to say that this one incompatible doctrine will spoil the party. But we might think that there needs to be a significant number of doctrines that support the conception. But how much is significant? My suggestion is to hypothesise that *most* publicly debated comprehensive doctrines must support a conception such that an overlapping consensus is possible.

2. It would be a very strong (and perhaps implausible) hypothesis to say: If it's not the case that most publicly debated doctrines support a conception, then an overlapping consensus is strictly speaking *impossible*. Instead, it makes more sense to speak in terms of probability instead of possibility. This yields as a hypothesis: If it's not the case that most publicly debated comprehensive doctrines support a conception, then it's *improbable* that there is an overlapping consensus on that concep-

tion.

(I will say more about these adjustments below.) Since I have distinguished different kinds of overlapping consensus and I am in this thesis only concerned with the potential kind, this gives us the following set of hypotheses. Let PC be a (publicly debated) political conception of justice.

Hypotheses L ('L' for 'local'). If it's not the case that most comprehensive doctrines in the (common core of the) dialectical situations support PC, then

1. it is improbable that there is a potential local overlapping consensus on PC in the weak sense.

2. it is improbable that there is a potential local overlapping consensus on PC in the strong sense.

3. it is improbable that there is a potential local overlapping consensus on PC of high grade, i.e. $r \geq 0.5$.

Hypotheses G ('G' for 'global'). If it's not the case that most comprehensive doctrines in the (common core of the) dialectical situations support PC, then

1. it is improbable that there is a potential global overlapping consensus on PC in the weak sense.

2. it is improbable that there is a potential global overlapping consensus on PC in the strong sense.

3. it is improbable that there is a potential global overlapping consensus on PC of high grade, i.e. $r \geq 0.5$.

A few remarks about these hypotheses.

First, the hypotheses may seem somewhat arbitrary. For example, why is it about *most* doctrines, i.e. more than 50%, and not, for example, more than 40% or 60%? Why is a local or global overlapping consensus of high grade defined as having a grade $r \geq 0.5$ and not, for example, $r \geq 0.7$ or $r \geq 0.8$? It is, indeed, somewhat arbitrary. The basic idea is just to weaken the implausibly strong hypothesis that *all* publicly debated doctrines must support a conception such that an overlapping consensus is *possible at all*.

But the arbitrariness of how exactly I weakened the hypotheses is not a problem, since they are mainly meant to guide the study design. When analysing the results, we can, in principle, switch things up and use different thresholds. But it will turn out that the data gives interesting answers about the hypotheses as they are stated here.

Second, the talk of probabilities has to be interpreted. This interpretation will, of course, heavily depend on the context. The context here is the set of artificial societies constructed in section 4. Roughly, these societies are randomly generated in a way that enables us to isolate the influence of the inferential connections between comprehensive doctrines and a given political conception. Statements about the probability of an overlapping consensus of some kind can then be made by analysing how many of the randomly generated societies with certain inferential connections exhibit that kind of overlapping consensus. Thus, I presuppose a frequentist interpretation of probability, as is usual in statistics and science in general (cf. Hájek, 2023). Of course, we are ultimately interested in real societies and the corresponding probabilities (and not artificial ones). In section 4.3, I explain what the artificial societies can tell us about real societies.

Third, it is not entirely clear what Rawls would say about these hypotheses. In sections 2.1.1–2.1.2 I discussed passages suggesting that Rawls thinks of the comprehensive doctrines in an overlapping consensus as *supporting* the political conception of justice. As a consequence, it seems plausible that Rawls would agree with the general idea behind the hypotheses, i.e. that the doctrines in the citizens' dialectical situations must support the conception such that an overlapping consensus is possible. However, there are other passages in which Rawls seems to endorse a lower standard. For example, when discussing how an overlapping consensus might come about, he speculates that even citizens with neutral or incompatible doctrines can come to accept the respective political conception of justice (PL 160). But there he is concerned with genesis and not justification. Rawls seems to imagine that citizens in the process of forming an overlapping consensus will adjust their doctrines such that they support the conception. On the bottom line, however, I think it is unclear whether testing the research hypotheses will confirm or disconfirm a view held by Rawls.

Finally, I think it is nonetheless important to find out whether these

hypotheses are true. Let's again set aside the complexities introduced above and remember the original conjecture that only support connections will foster an overlapping consensus, while incompatibility and neutrality are potentially threatening. If this turned out to be true (or not false), i.e. the predictions of the corresponding hypotheses are verified (or not falsified), then this is a worrisome outcome. It sets a high standard for a public debate that is compatible with realising an overlapping consensus. In particular, we might have to change how to conduct a public debate. In the most extreme form, this could mean completely excluding certain doctrines. For illiberal or anti-democratic doctrines (i.e. ones that are incompatible with a liberal democratic political conception), this might not seem as grave a problem. Maybe (just maybe!) one can argue that these are deeply and objectively wrong doctrines. But neutral doctrines are simply ones that do not take a stand on the political conception. If we had to exclude these or somehow treat them in a significantly different way than supportive ones, then this seems unfair. It goes even more strongly (than excluding incompatible doctrines) against the ideal of a public debate that is open to all views and arguments. Thus, in a nutshell, liberal democrats who are, such as me, in favor of an overlapping consensus account of societal stability and in favor of a liberal public debate open to all, wouldn't like if these hypotheses held. Testing them, e.g. as I do in this thesis, is crucial. It will give us a better idea of how to organise societies such that they can be both stable and liberal.

2.3 Summary

In this chapter I have thus far presented the main philosophical commitments of this thesis. I have done so mostly in discussion of the relevant passages in Rawls's *Political Liberalism* and *A Theory of Justice*. I have cleared up to which parts of the Rawlsian view I am committed, regarding which parts I am neutral and from which parts I deviate. Here is a summary of the results:

- I subscribe to the general idea of an *overlapping consensus*: The different comprehensive doctrines held by the citizens overlap on a shared political conception of justice. However, I am non-committal regarding almost all of the Rawlsian specifics: Where exactly to draw the

line between comprehensive doctrines and political conception, how to make a conception of justice freestanding, etc. (See section 2.1.1.)

- I subscribe to the idea that citizens in an overlapping consensus need to be *morally justified* in holding their beliefs. This is the gold standard of societal stability. (See section 2.1.2.)

- I presented semi-formal definitions of *different kinds of overlapping consensus*. The most important distinctions are: actual vs. potential, global vs. local, and, regarding the potential kind, weak sense vs. strong sense vs. graded sense. The ultimate goal is to have an actual global overlapping consensus, but the other kinds correspond to different intermediary stages. (See section 2.1.3.)

- Like Rawls and many other philosophers, I subscribe to *equilibrationism*, i.e. the idea that (moral) beliefs are justified by an equilibration process that makes the belief system as a whole coherent. I reject the idea that such equilibration processes must use the Rawlsian expository device of the initial situation. I am non-committal about what the appropriate starting point for equilibration processes is, but I do subscribe to the Rawlsian account of coherence consisting of derivability and systematicity. (See section 2.2.1.)

- I highlighted that the relevant kind of justification is what Rawls calls 'full justification' and it belongs to the *non-public sphere*. In particular, I am not concerned with public reason and public justification which depend on this full justification as a precondition. (See ibid.)

- Like Rawls, I subscribe to *Reconstructionism*, i.e. the idea that citizens need not actually go through an equilibration process in order to be justified. It is enough if their beliefs could have been the result of such a process. A consequence of this is that the present thesis is concerned with propositional, not doxastic justification. (See section 2.2.3.)

- I subscribe to *Epistemic Consequentialism* and embrace a *bounded rationality* perspective. This means that agents, in order to be justified, need to use a method that is both feasible for them and effective at making

their beliefs coherent, or their beliefs can be reconstructed to be the result of such a method. (See section 2.2.4.)

- I reject Rawls's conception of wide reflective equilibrium as too demanding. However, I do agree that we need some plausible idea regarding the citizens' *dialectical situation*, i.e. the views and arguments they must consider during equilibration. I suggested that the dialectical situation of a citizen is comprised of at least the views and arguments that are publicly debated in their society. (See section 2.2.5.)

When I presented the definitions for the different kinds of overlapping consensus in section 2.1.3, I left open what the set of justified belief systems for an agent is. Given the philosophical commitments from section 2.2, we can now give a definition:

Definition 5 (Propositional Justification: equilibrationist, reconstructionist, consequentialist, non-ideal)**.** Let B be the set of all possible (moral) belief systems. Let a be an agent in dialectical situation D with initial commitments $C_0^a \in B$. Then the set $J \subset B$ of *belief systems propositionally justified for* a is defined as: $b \in J$ iff b could have been the result of applying a feasible and effective equilibration method starting from C_0^a and considering D.

In chapter 3, I will give a formal explication of this definition and also several explications for the different kinds of overlapping consensus.

Finally, following my discussion of the citizens' dialectical situations, I formulated a detailed research question: 'Which kinds of inferential connections between publicly debated comprehensive doctrines and a (publicly debated) political conception make a potential overlapping consensus possible?'. This question is important but underresearched. I formulated several research hypotheses as potential answers to this question, one for each kind of overlapping consensus I distinguished earlier. The hypotheses state that most publicly debated comprehensive doctrines must support a political conception such that an overlapping consensus of a particular type can be probable. I argued that testing these hypotheses should be of great interest to liberal democrats. One central goal of the simulation study presented later is to test these hypotheses.

Chapter 3

Formal explications

The goal of this chapter is to turn the semi-formal definitions of the last section into formal explications of

- the notion of justification (sections 3.1–3.3)
- the notion of consensus (section 3.4),
- the notion of pluralism (section 3.4),
- and, putting these three together, the different notions of an overlapping consensus (section 3.5).

The biggest task by far is the first: explicating the notion of justification. So let's get to it.

3.1 The theory of dialectical structures

Beisbart, Betz and Brun (2021) present a formal model of MRE. The next two sections give a short introduction to the model. It is based on the theory of dialectical structures by Betz (2021). In this section, I explain the fundamentals of this theory.

A dialectical structure is supposed to represent the state of a debate concerning a certain subject matter. Formally, each such structure is an ordered pair of two sets:

- A *sentence pool* S, representing the subject matter. This set of sentences is closed under negation (with $\neg\neg s := s$).

- A set of *arguments A on S*, representing the deductive relations between the sentences. Each argument is an ordered pair of a set of premises from S and a conclusion from S.

Any subset of S is a *position*. This subset represents the sentences that the agent accepts. To reject a sentence means to accept its negation. Here is an example for such a structure and some positions on it:

$$S = \{ s_1, s_2, s_3, \neg s_1, \neg s_2, \neg s_3 \}$$
$$A = \{ (\{ s_1 \}, \neg s_3), (\{ s_2 \}, s_3) \}$$

$$P_1 = \{ \neg s_1 \}$$
$$P_2 = \{ s_2, s_3, \neg s_3 \}$$
$$P_3 = \{ s_1, s_2, s_3 \}$$
$$P_4 = \{ \neg s_1, s_2, s_3 \}$$

There are at least two salient differences between these positions. First, P_1 and P_2 are non-committal or neutral about some of the sentences, but P_3 and P_4 take a stand on all of them. We say that a position is *complete* iff it contains each sentence or its negation or both, otherwise it is *partial*. Second, P_2 contains both s_3 and $\neg s_3$. It contradicts itself in this obvious way while the other positions do not. We say that a position is *minimally consistent* iff it does not contain both a sentence and its negation.

Of course, the notion of minimal consistency is very basic. It does not take into account the inferential relations represented by the set of arguments A. For example, P_3 accepts both s_1 and s_3, but s_1 implies $\neg s_3$ according to one of the arguments. In that sense, P_3 is not consistent. This more robust notion is called *dialectical consistency* or simply *consistency*. We define this notion separately for complete and partial positions. A complete position is *(dialectically) consistent* iff

1. the position is minimally consistent and,

2. for every argument, if the position contains all premises, then it contains the conclusion.

The consistency of partial positions is defined with reference to the consistency of complete positions: A partial position is (dialectically) consistent iff the position is extended by some complete consistent position, i.e. is a subset of it. Finally, a position is *inconsistent* iff it is not consistent. What about our example? Of the complete positions, P_3 is inconsistent and P_4 is consistent.

Of the partial positions, P_2 is inconsistent (because it is not minimally consistent) and P_1 is consistent (because it is extended by P_4 which is a complete consistent position).

Lastly, let me introduce the notion of the content of a position. In this context, the content of a position is what the position deductively implies. For example, consider the consistent partial position $P_5 := \{ \neg s_3 \}$. Intuitively, P_5 implies $\neg s_3$ and $\neg s_2$: It implies $\neg s_3$ trivially and $\neg s_2$ by contraposition of one of the arguments. Formally, the content of a consistent position P is represented by the intersection of all consistent complete positions that extend P. This intersection is again a position (denoted $\overline{P}$) and contains all sentences that the original position implies. In the case of P_5, there are two consistent complete positions that extend it, $\{ s_1, \neg s_2, \neg s_3 \}$ and $\{ \neg s_1, \neg s_2, \neg s_3 \}$. The intersection of these is $\overline{P_5} = \{ \neg s_2, \neg s_3 \}$, matching the above intuition.

This concludes my introduction to the theory of dialectical structures. Before I go on, however, I wish to make two important comments. First, let me stress how much hinges on *adequately* representing a dialectical situation. This includes the logical relationships between the sentences in S, because they need not be atomic. For example, s_2 could represent 'The sun is bright' while s_3 represents 'The sun is bright and it is warm'. In this case, the above set of arguments A is an *inadequate* representation of the deductive relations between the sentences, because it says that s_3 follows from s_2 even though it is the other way around. This potential mismatch is the price for the liberty one has when modelling a dialectical situation. The advantage, on the other hand, is that the theory of dialectical structures is compatible with many different systems of logic. In what follows, I always assume that there is an interpretation of the sentences in S such that the arguments A are an adequate representation of the deductive relations between the sentences. Also, I always assume that the structures are *satisfiable* (there is at least one consistent complete position), e.g. liar paradoxes are excluded.

Second, even though the set of arguments A are supposed to be *deductive* arguments, they can be also interpreted as presupposing *uncontroversial supporting premises*. For example, suppose that s_2 represents 'Donald Trump is orange' and s_3 represents 'There are orange humans'. If we are modelling a dialectical situation in which it is uncontroversial that Donald Trump is

a human (and not, for example, a lizard person or a pseudo-intelligent robot deployed by the deep state), then we need not explicitly represent this supporting premise in the structure and can instead just add the argument ($\{ s_2 \}$, s_3) to the structure. This point will recur in section 4.1.

3.2 A model of reflective equilibrium

Based on these notions from the theory of dialectical structures, Beisbart, Betz and Brun (2021) present a formal model of the method of reflective equilibrium. The dialectical structure (S, A) is assumed to be given and fixed. It represents the agent's dialectical situation, i.e. the views and arguments they need to consider during equilibration. At any time, the agent's epistemic state can be represented by a pair of positions (C, T) where C is called the *commitments* of the agent (has to be minimally consistent) and T is called the *theory* of the agent (has to be consistent). The sentences in T are called the theory's *principles*. This formal notion of an epistemic state explicates the informal notion of a belief system from the last chapter:

Explication 1 (Belief system)**.** Let (S, A) be a dialectical structure. Then the *belief system* b of an agent is explicated as

$$b := (C, T),$$

with the minimally consistent *commitments* of the agent $C \subset S$ and the consistent *theory* that the agent accepts $T \subset S$.

There is, of course, an open question of interpretation concerning what it means for an agent to be committed to a set of sentences or to accept a set of sentences as their theory. I will leave this open, since the results of the present thesis will not depend on it, as long as there is some interpretation such that the explications presented in this chapter are plausible. If you need some idea, then you can take the commitments of the agent to be their *beliefs* about the subject matter S (cf. Daniels, 1979). The theory, on the other hand, you can take to be what the agent uses to give a *systematic account* of the subject matter S (cf. Baumberger and Brun, 2021, sec. 2.3).

This pair of positions plainly mirrors the two components in Rawls's

presentation of MRE: the commitments correspond to the considered judgments, the theory corresponds to, well, the theory, and the theory is supposed to somehow match and account for the commitments. Concerning the equilibration process, the idea is again similar to Rawls: starting from some initial commitments C_0, we go back and forth between theory and commitments and make adjustments that improve the epistemic state until no further improvement is possible.

To make sense of this idea of improvement, the model defines an *achievement function*:

$$Z(C, T|C_0) \;=\; \alpha_A \cdot A(C, T) \;+\; \alpha_S \cdot S(T) \;+\; \alpha_F \cdot F(C|C_0)$$

This real-valued function is the sum of three inner functions with non-negative weights $\alpha_A + \alpha_S + \alpha_F = 1$. The inner functions are called:

- *Account*, $A(C, T)$: measures how close the current commitments C are to the content $\overline{T}$ of the current theory T. This is supposed to reflect how well the theory matches the commitments and accounts for them.
- *Systematicity*, $S(T)$: measures how systematic the current theory T is. Less principles and more content improve this function.
- *Faithfulness*, $F(C|C_0)$: measures how close the current commitments C are to the initial ones C_0. The idea here is to have some tie to the starting point such that an agent cannot without good reasons discard the commitments they started with.

For the mathematical definitions, see the appendix. These functions represent desiderata for epistemic states (Beisbart et al., 2021). Regarding account, this is obvious. It is the central goal of MRE to adjust commitments and theory to each other until they match. However, it is plausible that this is not the only desideratum, because it would be too easy to come by: Just pick your initial commitments as your theory ($T := C_0$) and be done with it. This can't be it. Such a theory is hardly a theory in any substantial sense, at least in most cases. In particular, it is usually not very *systematic*. Systematic theories are simple to state, yet rich in content. A paradigm of a systematic moral theory is utilitarianism: a single principle that entails the moral

status of every conceivable action in every conceivable world. Of course, also less systematic theories are permissible, but systematicity is nonetheless a desideratum.

However, we still need more than account and systematicity. Otherwise, we would simply have to choose the most systematic theory T (or any of them if there are several) and pick its content as our commitments ($C := \overline{T}$). We would have to do this no matter how outlandish or counterintuitive the theory is. Surely this can't be right. We might even worry that we are simply changing the subject in accepting a theory that goes against all intuition and common sense. Of course, the most systematic theory need not be very outlandish, but the mere possibility of this worst-case-scenario shows that something is missing. Thus, we need a third desideratum, faithfulness, which establishes a tie to the initial commitments. The purpose of this tie is to ensure that we cannot without good reasons discard the initial commitments. Of course, we still can and should depart from them if it is worth it in terms of account and systematicity. (In general, 'good reasons' can be many more things, of course, but in this model it is a gain in account or systematicity.) Still, this tie ensures that we don't change the subject (for more on this point, see Baumberger and Brun, 2021). Moreover, to the extent that the initial commitments have a strong enough epistemic standing, it fends of the objection to pure coherentism that *any* coherent belief system, even the most absurd, counts as justified (see section 2.2.1). It is this tie that makes equilibrationism a weakly foundationalist, instead of coherentist, account of justification.

Now, the weights $\alpha_A, \alpha_S, \alpha_F$ in the achievement function make the trade-off between these desiderata explicit. As a standard configuration, $\alpha_A = 0.35$, $\alpha_S = 0.55$, $\alpha_F = 0.1$ has proven to yield plausible results. For a more detailed exposition and motivation of the achievement function, see Beisbart, Betz and Brun's 2021 paper.

Note that the desiderata of account and systematicity very straightforwardly explicate the Rawlsian requirements of derivability and systematicity (section 2.2.1), i.e. the idea that in a coherent belief system, the considered judgments must be derivable from systematic principles. This is why I said that I am substantially committed to these ideas. Together, account and systematicity are, for the purposes of the present model, an explication of

the notion of *coherence*. (Alternative ways of explicating coherence include (Tersman, 1993; Thagard, 2000), see section 6.3.) The desideratum of faithfulness is added to coherence, ensuring a tie to the initial commitments for the reasons mentioned above. All three components (account, systematicity and faithfulness) are merged together in the achievement function. Thus, the achievement function explicates the degree to which a belief system is in the state of reflective equilibrium (short: its degree of equilibrium), see section 2.2.4.

Also note that for account and faithfulness the achievement function measures distances: For account, it measures the distance between the commitments and the theory's content. For faithfulness, it measures the distance between the commitments and the initial commitments. Both distances are variants of the so-called Hamming distance, normalised to return values between 0 and 1 (see appendix A). However, both account and faithfulness are defined as the *closeness* of the two positions, not their distance. Thus, we need to use a monotonically decreasing function to transform these distances into appropriate measures for account and faithfulness. In fact, given the specific way that systematicity is defined, we also need such a monotonically decreasing function here (again, see the appendix for the details). Of course, there are infinitely many such functions. Beisbart et al. (2021) use $G_{quadratic}(x) := 1 - x^2$, but the model has been tested also for $G_{linear} := 1 - x$. The simulation study presented in the later chapters was conducted using both of these functions in order to see whether the results are robust regarding this point. Thus, keep in mind that, strictly speaking, there are two versions of the achievement function, though I will often just talk of *the* achievement function.

I wish to stress once more that the achievement function does not aim for 'mere consistency' or a 'mere match' between theory and commitments. Instead, it aims for *coherence*. The desideratum of systematicity urges the agent to choose a theory that *systematises* the commitments and in this sense explains them, makes sense of them, or helps us understand them. Importantly, the relation of explaining or making sense is not a relation between sentences. (The structure itself contains only *inferential* relations.) Instead, this relation is more of a macro feature that obtains between two positions, namely a highly systematic position (the theory) the content of which matches an-

other position (the commitments).

Now that we have explicated degrees of equilibrium, which is of sole epistemic value (see Epistemic Consequentialism in section 2.2.4), let's turn to the question of how to improve the epistemic value of one's epistemic state, i.e. the *algorithm* of MRE. Let dialectical structure and initial commitments C_0 be given. Now, out of all consistent positions on the structure, a theory T_1 is chosen that maximises the achievement function for the initial commitments, i.e. maximises $Z(C_0, T|C_0)$. If two or more score best, we make a random choice between those. Then, we adjust the commitments: Out of all minimally consistent positions, a new set of commitments C_1 is chosen that maximises the achievement function for the current theory, i.e. maximises $Z(C, T_1|C_0)$ (again we make a random choice in case of a draw). We then go back and forth, holding the commitments (or theory) fixed while maximising achievement by chosing a new theory (or new commitments). This goes on until no adjustment of either theory or commitments improves achievement anymore: we have reached an *equilibration fixpoint*.

The definitions of achievement function and algorithm together yield a precise specification of the method of reflective equilibrium. Given a subject matter (i.e. dialectical structure) and initial commitments (i.e. a position), it says how exactly adjustments are made and when an end state (i.e. fixpoint) is reached. In fact, the model can be programmed such that computers can run equilibration processes. Of course, this model is not the only plausible specification of MRE. For one thing, one can also model MRE on completely different assumptions (e.g. Freivogel, 2021; Baumgaertner and Lassiter, 2023; Dellsén, 2024), see section 6.3. For another, even when staying close to the above assumptions, the present model is just one plausible explication of MRE. For example, it is possible to conceive of more or different desiderata (perhaps with a focus on 'epistemic virtues', see Freivogel, 2023b). Also, the mathematical definitions of these functions, particularly the specific values of the various weights in them, are to some extent arbitrary (as is to be expected). Last but not least, the algorithm may be changed to be more effective or more feasible or both (see below). Thus, it is possible that both the achievement function and algorithm will be superseded by more plausible or more elaborate versions.

However, as a deliberately simple starting point, the model seems plaus-

ible enough, especially since it has undergone quite thorough testing. In their 2021 paper, Beisbart, Betz and Brun discuss some equilibration processes on a specific example structure. They argue that the results are plausible and match our pre-theoretic expectations of MRE (2021, sec. 3). Also, they prove some basic analytic results that lend further plausibility to the model (2021, sec. 2.4). In addition, the research group around Beisbart, Betz and Brun, of which I am a member, has assessed the model by running simulations on large sets of randomly generated structures and analysing the results. The findings are publicly available in a recent technical report by Freivogel and Cacean (2023). It seems to me that this analysis corroborates the model, at least for the large part. This is not to say that it's all done and dusted. But it warrants treating the model as a starting point for further research: applying it to interesting scenarios, examining variations and extensions, etc.

In fact, one such variation is better suited for the present study than the original model, as I argue next.

3.3 Changing the model: Local optimisation

Remember from section 2.2.4 that I embrace both epistemic consequentialism and a bounded rationality perspective. That is, first, I interpret the method of reflective equilibrium as a mere means to an end, namely increasing achievement. Second, I suppose that this method has to be feasible for the agents. In this section, I present an alternative algorithm that is feasible for real agents and nonetheless relatively effective at reaching fixpoints with high achievement, at least under certain circumstances.

Given that the algorithm is just a means to an end, it is clear that we are not strictly bound to a particular version of it. For example, instead of using the semi-global algorithm from the last section, we could opt for a *global* one: Calculate achievement for *all* combinations of commitments and theory and choose (one of) the best one(s). Or, quite the contrary, we could stick to the step-wise adjustment of the semi-global algorithm, but optimise *locally* by looking for the best commitments in the close neighborhood of the previous commitments, likewise when adjusting the theory. This kind of piece-meal change is most likely what Goodman, one of the earliest proponents of MRE, had in mind (1955, p. 67). Since I am going to opt for this kind of local

algorithm, let me give you a detailed definition.

The basic idea is to change only single sentences. That is, the set of candidate commitments in any adjustment step contains all minimally consistent positions that *either*

- extend the current commitments by any one sentence (negations included), *or*
- are extended by the current commitments by any one sentence (negations included), *or*
- result from removing any one sentence (negations included) from the current commitments and adding that sentence's negation (with $\neg\neg s := s$).

Now, we calculate the achievement of these candidate commitments and choose the best one (selecting at random in case of a draw). For adjusting the theory, we proceed exactly the same way except requiring *(dialectical) consistency* instead of only minimal consistency. Note that we start out with a set of independently given initial commitments, so we can construct the first set of candidate commitments from these. We do not have such an independently given first theory. Thus, we must define some such theory and do so by setting $T_0 := \emptyset$. There are other ways of defining T_0 (e.g. Flick, 2022), but in the interest of keeping it simple (and feasible), the empty set seems a plausible enough starting point. Let's call this algorithm *LocalQuadraticMRE* or *LocalLinearMRE*, depending on whether we are using $G_{quadratic}$ or G_{linear} in the definition of the achievement function.

For my purposes, this local algorithm is more suitable than the semi-global or global one. This is because it is much more feasible, yet at the same time pretty effective for increasing achievement. Regarding effectiveness: Flick (2022) has tested the local algorithm and compared it to the semi-global one. His general upshot is that the local algorithm is as good as the semi-global one. However, this result only holds in structures with one-premise-arguments. This is one of the reasons why the study design (section 4.1) features only one-premise-arguments. (Presumably, the local alogrithm would have to consider changes to more than one sentence per step in order to work for arguments with more than one premise.)

Regarding feasibility: The local algorithm is *much* more feasible than the more global versions. To see this, consider an unrealistically small sentence pool of size $|S| = 40$ (including negations). This already yields $3^{20} \approx 3.5$ billion minimally consistent positions (i.e. commitments candidates). The semi-global algorithm requires going through all of them and calculating which most improves achievement. This requires *crazy* computational resources. Even the most advanced computers fail at this, let alone human brains. This makes it unsuitable for the present research. Not only, because the simulations could not be run (in my study, $42 \leq |S| \leq 54$). More importantly, as I already pointed out in section 2.2.4, for matters of stability we wish to find *realistic* conditions for an overlapping consensus, in particular, conditions that epistemically non-ideal citizens can satisfy. Not much is gained if we find conditions that can only be satisfied by currently unavailable supercomputers with extreme computational power. Instead, I try to stay closer to the cognitive capacities of actual citizens. The local algorithm is suitable for this aim. In the above example, we now only have 40 instead of 3.5 billion commitments candidates in each step. To be sure, it is still a lot of work for a human brain and it is an open question how feasible it really is. Nonetheless, it is a simple and initially plausible idea for an algorithm that has at least a chance of being feasible for us. For these reasons, I think the local algorithm is the right choice for researching conditions for overlapping consensuses in liberal democracies.

Before connecting the formal apparatus to the research question in a more detailed manner, let me recap. Beisbart, Betz and Brun have put forward a formal model of reflective equilibrium. It consists of two parts, the achievement function and an algorithm. The model has been tested, with good results. However, the semi-globally optimising algorithm is computationally demanding. In a bounded-rationality setting, which I am embracing, a locally optimising algorithm is the better choice. The local algorithm has also been tested with promising results. I think that the current status of the model warrants application to interesting cases like a overlapping consensus.

Given the results of the last few sections, we are now in a position to give an explication of the notion of justification that was defined in the last chapter. This was the final definition (section 2.3):

Definition 5 (Propositional Justification: equilibrationist, reconstructionist,

consequentialist, non-ideal)**.** Let B be the set of all possible (moral) belief systems. Let a be an agent in dialectical situation D with initial commitments $C_0^a \in B$. Then the set $J \subset B$ of *belief systems propositionally justified for a* is defined as: $b \in J$ iff b could have been the result of applying a feasible and effective equilibration method starting from C_0^a and considering D.

We can now give two explications of this notion, one for the quadratic achievement function and one for the linear one (they presuppose explication 1 of belief systems given above):

Explication 2 (Propositional Justification: equilibrationist, reconstructionist, consequentialist, non-ideal)**.** Let a be an agent with dialectical structure (S, A). Let $\mathcal{E} = \mathcal{C} \times \mathcal{T}$ be the set of possible epistemic states with the set of all minimally consistent positions $\mathcal{C} \subset \wp(S)$ (the possible commitments) and the set of all dialectically consistent positions $\mathcal{T} \subset \wp(S)$ (the possible theories). Let $F((S, A), C_0, Alg) \subset \mathcal{E}$ be the set of all possible fixed points (in particular, considering all random choices) of algorithm Alg applied to initial commitments C_0 on dialectical structure (S, A). Let $C_0^a \in \mathcal{C}$ be a's initial commitments. Then the set $\mathcal{J}^a \subset \mathcal{E}$ of *epistemic states propositionally justified for a* is explicated as:

$$\mathcal{J}^a_{quadratic} := F((S, A), C_0^a, LocalQuadraticMRE), \text{ or}$$
$$\mathcal{J}^a_{linear} := F((S, A), C_0^a, LocalLinearMRE)$$

These explications of the notion of justification are the basis for the simulation study presented later. They are, as I have emphasised already, just two possible such explications. Thus, it remains to be seen in how far the results are robust when we consider alternative explications (for an overview of these, see section 6.3).

Before we go on to explicating pluralism and consensus, a brief remark on epistemic permissiveness: The local algorithm (just like the semi-global one) prescribes random choices when more than one of the adjustment options maximise achievement. As a consequence, for each starting point (a set of initial commitments and a dialectical structure) there may be several fixpoints that can be reached. (This is the reason why I have distinguished different senses for the potential kinds of overlapping consensus in section

2.1.3.) A direct consequence of this is that this explication of justification leads to a certain *intra-personal epistemic permissiveness*, meaning that given the epistemic context of an agent (here dialectical structure and initial commitments) there is more than one epistemic state that is justified for the agent. This seems to conflict with the evidential uniqueness thesis stating that given a total body of evidence there is at most one rational or justified epistemic state regarding any proposition, or some version of this claim (for an excellent overview of different versions of evidential uniqueness, see Briesen, 2017). At least, it might conflict if we can plausibly cash out 'total body of evidence' in terms of dialectical structure and initial commitments. I will not here discuss whether this is a problem for the present explication. But note that it is not really a surprising outcome (Rawls himself considers this possibility in PL 44; see also White, 2005, p. 446) and it certainly does not entail a problematic 'anything goes' relativism (see DePaul, 2013, p. 4474; and especially Freivogel, 2023b, ch. 11, who considers this objection for the present model).

3.4 Consensus and pluralism

The result of the last sections is an explication of the notion of justification. However, if we want to explicate the different notions of an overlapping consensus from the last chapter, we need more than that: We also need to give a formal charactarisation or explication of the notions of *consensus* and of *pluralism*. That is, given a tuple of justified belief systems (one belief system for each citizen) and a political conception of justice *PC*, we want a formal criterion for saying whether that tuple exhibits a pluralism of comprehensive doctrines and consensus on *PC*. The goal of this section is to present these explications. Of course, there are many ways of measuring consensus and pluralism. For some examples regarding consensus, see (Diamond et al., 2014; Holey et al., 2007; Alcalde-Unzu and Vorsatz, 2011), regarding pluralism, see (Haidt et al., 2003; Bramson et al., 2017; Singer et al., 2019; Osborne and Atari, 2024). My project, however, requires that the measures operate on tuples of epistemic states as explicated above and, more importantly, target specific parts of these epistemic states, namely, political conceptions and comprehensive doctrines. For these reasons, I develop my own meas-

ures for consensus and pluralism, custom-tailored to the investigation of overlapping consensus.

I should say right from the start that I will give explications of the *gradual* senses of consensus and pluralism, i.e. the senses in which a society has *more or less* consensus on *PC* and is *more or less* pluralist regarding the comprehensive doctrines. But the definitions of the different kinds of overlapping consensuses mention the *categorical* senses of these notions. In order to get an explication of these categorical senses, we would need to set definite *thresholds* for the gradual explications. This is not an easy task. But, luckily, in section 5.2 it will turn out that this is not necessary for the present purposes.

For the rest of this section, let $A = \{a_1, ..., a_n\}$ be a set of agents living together in a society and sharing the dialectical structure (S, A). Let $\mathcal{E} = \mathcal{C} \times \mathcal{T}$ be the set of possible epistemic states with the set of all minimally consistent positions $\mathcal{C} \subset \wp(S)$ (the possible commitments) and the set of all dialectically consistent positions $\mathcal{T} \subset \wp(S)$ (the possible theories). Let $\mathcal{J}^{a_i} \subset \mathcal{E}$ be the set of epistemic states justified for agent a_i and $\mathfrak{E} := \mathcal{J}^{a_1} \times \cdots \times \mathcal{J}^{a_n}$ the space of justified belief systems for their society. Let $PC \in S$ be a political conception of justice.

3.4.1 A measure of consensus

Let's start with consensus. The easiest and most straightforward way for explicating the extent to which citizens agree on a political conception *PC* is to calculate the *acceptance rate of PC* in a society:

Definition 6 (Acceptance rate)**.** The *acceptance rate of PC* in a tuple $((C_1, T_1), \ldots, (C_n, T_n)) \in \mathfrak{E}$ is given by:

$$AccRate((C_1, T_1), \ldots, (C_n, T_n)) = \frac{100}{n} |\{C_i : PC \in C_i\}|$$

This definition says that the acceptance rate of *PC* of a tuple of epistemic states is equal to the percentage of states accepting *PC* in the commitments. Thus, if half of the citizens are committed to *PC* in that tuple, then the acceptance rate in that tuple is 50. If all citizens are committed to *PC* in that tuple, then there is a maximal acceptance rate of 100. Note that it does not

matter whether *PC* is in the theory of an epistemic state or not.

Since this definition is simple and plausible, I will adopt it as an explication of the gradual notion of consensus. Again, to get an explication of the categorical notion, one would have to set a plausible threshold $t_{AccRate}$ (presumably above 50), but I here rest content with the gradual notion.

3.4.2 Three measures of pluralism

Let's turn to measuring the pluralism of comprehensive doctrines in a tuple of commitments, i.e. explicating the degree of this pluralism. In principle, for each tuple one can measure pluralism in the whole society as well as in parts of the society, i.e. in subsocieties. As we have seen in section 2.1.3, for the local kinds of overlapping consensuses on *PC* we are interested in the pluralism in the subsociety of agents accepting *PC*, i.e. in the *PC*-subsociety of a tuple. This particular kind of pluralism I will often just call *PC*-pluralism, or simply pluralism. The global kinds of overlapping consensuses were defined with reference to the 'global' pluralism in a society. However, as I explain later in this chapter, it is sensible to also explicate the global notions using *PC*-pluralism instead of global pluralism. Thus, since I am only concerned with *PC*-pluralism, I will often just talk of pluralism instead of *PC*-pluralism. Whenever I mean global pluralism in particular, I will explicitly say so. I use the variable n_{FP} to denote the number of agents in the relevant subsociety of a tuple. (The subscript 'FP' stands for 'fixpoints', because n_{FP} likewise denotes the length of the respective subtuple of fixpoints, see the definition below.)

In what follows, I present different measures for pluralism in any given subtuple of a tuple. But, as I just said, we are ultimately interested in a particular subtuple, namely the one containing all epistemic states accepting *PC*. Let's define it.

Definition 7 (Subtuple and *PC*-subtuple)**.** Let $\mathfrak{T} = ((C_1, T_1), \ldots, (C_n, T_n)) \in \mathfrak{E}$. Let $I \subset \{1, \ldots, n\}$ with cardinality $m := |I| = n_{FP}$ and elements $s_1 < ... < s_m$. Then $((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m}))$ is a *subtuple* of $\mathfrak{T}$. In particular, let $I_{PC} := \{i \in \{1, \ldots, n\} : PC \in C_i\}$ with cardinality $m := |I_{PC}|$ and elements $s_1 < ... < s_m$.

Then the *PC-subtuple* of $\mathfrak{T}$ is defined as

$$\mathfrak{T}^{PC} := ((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m})).$$

This definition mirrors definition 3 in section 2.1.3, only that it is about epistemic states instead of belief systems in general.

There are several ways of measuring pluralism, each focusing a different aspect of the notion. In this study, I employ three measures of pluralism: *option count*, *strength of the weak* and *entropy*. The measures work very differently and the numbers that the functions produce are hardly comparable. Nonetheless, the measures share two features.

First, all three measures operate on the *distribution of CD-options* in the *PC*-subsociety. Suppose there are a number of n_{CD} comprehensive doctrines in the dialectical structure. Thus, the comprehensive doctrines in the structure are $CD_1, \ldots, CD_{n_{CD}}$. Given that the doctrines are incompatible and, as a consequence, any fixpoint will contain at most one of these, there are $n_{CD} + 1$ possible CD-options that can be realised in any fixpoint: Opt_{CD} := { {}, {CD_1}, ..., {$CD_{n_{CD}}$} }. Any fixpoint can contain any one of the comprehensive doctrines or none of them. For example, suppose that $n_{FP} = 25$ agents accept *PC* in a given tuple and $n_{CD} = 4$. Then the distribution of CD-options in the *PC*-subsociety of the tuple could be: 2×{}, 5×{CD_1}, 12×{CD_2}, 6×{CD_3}, 0×{CD_4}. In fact, when normalised appropriately, any such distribution is a *probability distribution* over Opt_{CD} (see appendix B). For each measure, this distribution is all that is needed to calculate *PC*-pluralism.

The second feature that all measures share is that I normalised them such that the following conditions hold: If all agents in a subsociety realise the same CD-option, then pluralism is minimal and the function is 0. If all CD-options are realised the same number of times, i.e. there is a *homogeneous* distribution of CD-options, then pluralism is maximal and the function is 100. (I use the term 'homogeneous' as synonymous to 'uniform'.) In the above example, every CD-option would have to be realised 5 times. The term 'homogeneous' (like 'uniform') is, of course, misleading, because when thinking of a homogeneous society we think of a society with little pluralism. But here, being homogeneous is a property of the distribution, not society itself.

There is a property of subsocieties that comes in handy when normalising: The maximum number of realisable CD-options. Often, the maximum number of realisable options *max* will just be $max = |Opt_{CD}| = n_{CD} + 1$. However, if the number of agents n_{FP} in the subsociety of a tuple is smaller than $|Opt_{CD}|$, then $|Opt_{CD}|$ can never be reached and the maximum number of realisable options is instead $max = n_{FP}$, i.e. the case when every agent in the subsociety realises a different CD-option. Therefore, the normalising factor I will use (in a different way for each measure) is:

$$max := \min(\{\, n_{CD} + 1,\ n_{FP} \,\}).$$

Now, let's look at the different measures for pluralism.

Option count

I think it's best to start with the most minimal notion of pluralism. According to *option count*, the only thing that matters for pluralism is the number of CD-options that are realised at least once. This is, of course, a feature of the distribution of CD-options as defined above, but option count is not at all sensitive to how the agents of a subsociety of a given tuple are distributed over the CD-options that are realised at least once. In this sense, option count is *distribution-insensitive*. That is, according to option count, even if there is one very dominant option (with a lot of agents realising it) and several small minorities, the mere fact that there *are* minorities tells us that the society is pluralist. On this view, pluralism means that many of the possible CD-options are not excluded and have some place, however small, in the subsociety of the tuple.

Definition 8 (Option count)**.** The *option count* of a subtuple $((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m}))$ (with $m = n_{FP} \leq n$) of some tuple $((C_1, T_1), \ldots, (C_n, T_n)) \in \mathfrak{E}$ is given by

$$OptCount((C_{s_1}, T_{s_1}), \ldots, (C_{s_n}, T_{s_n})) := 100 \cdot \frac{|\{O \in Opt_{CD} : \exists C_{s_i} \text{ s.t. } C_{s_i} \text{ realises } O\}| - 1}{max - 1}.$$

Roughly speaking, this definition counts the number of realised options and

normalises it with the maximum number of realised options. Additionally, it subtracts 1 from the numerator, because we want the function to be 0 if all agents realise the same option. In other words, one realised option comes free and does not count towards pluralism. It subtracts 1 from the denominator so that the function is 100 when the maximum number of realisable options is in fact realised. Strictly speaking, option count measures the ratio of the number of *realised* options (beyond 1) to the number of *realisable* options (beyond 1).

Strength of the weak

Let's turn to *strength of the weak*, the second measure of pluralism. For option count it does not matter whether there is a very dominant CD-option as long as a significant number of the remaining CD-options is realised at least once. Strength of the weak, on the other hand, is the opposite in that regard. The basic idea behind this measure is that dominance is the enemy of pluralism. If the strongest option is very strong, i.e. a lot of fixed point commitments realise this option, then the society is less pluralist. In other words, if 'the weak', i.e. all agents that do not realise the strongest option, are comparatively strong, then the society is comparatively pluralist: Pluralism is strength of the weak. This idea has some initial plausibility, of course.

Definition 9 (Strength of the weak)**.** Let $((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m}))$ (with $m = n_{FP} \leq n$) be a subtuple of $((C_1, T_1), \ldots, (C_n, T_n)) \in \mathfrak{E}$. Let $O_{strong} \in Opt_{CD}$ be the strongest CD-option (or one of them if there are several), i.e. the CD-option that is realised in the commitments of at least as many epistemic states in the subtuple as any other CD-option. The *strength of the weak* of $((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m}))$ is given by

$$SoW((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m})) := 100 \cdot \frac{|\{C_{s_i} : C_{s_i} \text{ does not realise } O_{strong}\}|/n_{FP}}{(max - 1)/max}.$$

The function takes the proportion of the weak to the whole subsociety (the numerator). It is then normalised with the highest possible such proportion (the denominator): the case of a homogeneous distribution of agents over all options. In this case, the weak are maximally strong, e.g. for five realisable

options ($max = 5$) they have a strength of $(max - 1)/max = 80\%$. Thus, strength of the weak measures the actual strength of the weak (numerator) in relation to the theoretically possible strength of the weak (denominator).

Note that it follows that strength of the weak is *distribution-sensitive* in a way that option count is not. For option count the distribution is *only* relevant to the extent that it gives us the options that are realised at least once, but the distribution of *these* options is not at all important. Strength of the weak, on the other hand, is sensitive to this distribution, but only regarding the distribution of agents over the strongest vs. the other options. It does not matter for strength of the weak how the agents are distributed over these other options.

Entropy

Finally, *entropy* is a generalisation of the idea behind strength of the weak. Like strength of the weak, it takes dominance to be the enemy of pluralism. But while strength of the weak is only sensitive to the distribution of agents over the strongest vs. the other options, the distribution-sensitivity of entropy holds for *any* part of the distribution. In particular, if there is a strongest option with, say, 60% of the fixpoints realising it, then entropy will still be sensitive to how the remaining 40% of the agents are distributed over the weaker options. If there is some option that most of those 40% agents realise, then the overall entropy will be lower than if those 40% are more evenly spread over those options. (This follows from the so-called additivity of entropy, see Aczél et al., 1974, for a definition.) Holding any part of the distribution as fixed, 'dominance is the enemy of pluralism' applies to the remaining distribution.

In essence, entropy is a measure for how homogeneous a probability distribution is, in this case, for how homogeneous the distribution of CD-options is. Thus, it is maximal if and only if every CD-option is realised exactly the same number of times. And this also applies to parts of the distribution when holding the rest fixed. Any deviation from a perfectly homogeneous distribution is seen as a form of dominance which is the enemy of pluralism. A maximally homogeneous distribution corresponds to a maximally pluralist society.

This interpretation of entropy as a measure for the homogeneity of a distribution is underpinned by its connection to the Kullback-Leibler divergence (short: KL divergence), originally introduced by Kullback and Leibler (1951). KL divergence has become a standard measure for how *different* a probability distribution is from some other distribution. In particular, we can use it to measure how different a distribution is from the homogeneous distribution. If we multiply this measure by -1, we get a measure for how *similar* a distribution is to the homogeneous distribution. When normalised appropriately, the resulting measure is equivalent to the entropy of the distribution (see appendix B for a proof).

Thus, entropy is a measure for the homogeneity of a distribution which in turn is a measure for the pluralism in a society, if we completely follow the idea that dominance is the enemy of pluralism. Here is a definition (cf. Shannon, 1998, though I adapt to the present circumstances):

Definition 10 (Entropy)**.** Let $((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m}))$ (with $m = n_{FP} \leq n$) be a subtuple of $((C_1, T_1), \ldots, (C_n, T_n)) \in \mathfrak{E}$. For $O \in Opt_{CD}$ let $p(O) := |\{C_{s_i} : C_{s_i} \text{ realises } O\}|/n_{FP}$. The *entropy* of $((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m}))$ is given by

$$Entropy((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m})) := 100 \cdot (-1) \sum_{O \in Opt} p(O) \log_{max} p(O).$$

Choosing *max* as the base for the logarithm ensures that the entropy is 100 if the distribution of the CD-options is homogeneous just like for the other measures.

This finishes my presentation of the three measures of pluralism which can serve as the explications of the gradual notion of pluralism in a subsociety of a tuple of epistemic states. I have presented these measures in their order of increasing distribution-sensitivity.

3.5 Explicating kinds of overlapping consensus

Given the explication of justification (explication 2) from section 3.3 and the definitions for measures of consensus and pluralism from the last section, we are now in a position to give explications for the different notions of an overlapping consensus defined in the last chapter. Let's first restate these

definitions:

Let $A = \{a_1, ..., a_n\}$ be a set of agents living together in a society. Let B be the set of all possible belief systems. Let $J_i \subset B$ denote the finite set of belief systems that are justified for agent a_i. Let PC be a political conception of justice.

Definition 2 (Potential Global Overlapping Consensus)**.** There is a *potential global overlapping consensus on PC*

- *in the weak sense* iff there is at least one tuple from $J_1 \times ... \times J_n$ that exhibits a pluralism of comprehensive doctrines and a consensus on PC.
- *in the strong sense* iff all tuples from $J_1 \times ... \times J_n$ exhibit a pluralism of comprehensive doctrines and a consensus on PC.
- *of grade r* iff a proportion $r \in [0, 1]$ of all tuples from $J_1 \times ... \times J_n$ exhibits a pluralism of comprehensive doctrines and a consensus on PC.

Definition 4 (Potential Local Overlapping Consensus)**.** There is a *potential local overlapping consensus on PC*

- *in the weak sense* iff there is at least one tuple from $J_1 \times ... \times J_n$ that exhibits a pluralism of comprehensive doctrines in its PC-subsociety.
- *in the strong sense* iff all tuples from $J_1 \times ... \times J_n$ exhibit a pluralism of comprehensive doctrines in their respective PC-subsocieties.
- *of grade r* iff a proportion $r \in [0, 1]$ of all tuples from $J_1 \times ... \times J_n$ exhibits a pluralism of comprehensive doctrines in their respective PC-subsocieties.

Note that the different senses of a potential global overlapping consensus do not require the respective tuples to exhibit a pluralism of comprehensive doctrines in their respective PC-subsocieties (like the local kinds), but throughout the tuples. That is, they require *global* pluralism in each tuple, as I have already mentioned at the beginning of the last section. Technically speaking, even though the different pluralism measures are defined for subtuples, they can nonetheless be used for measuring global pluralism, because the definition of a subtuple is such that a tuple counts as a subtuple of itself

(similar to a set being a subset of itself). Nonetheless, I will not explicate the senses of a potential global overlapping consensus by measuring global pluralism. Here's why: There may be consensus on *PC* in a tuple of belief systems even though not all of them agree on *PC*. In other words, when setting a threshold for acceptance rates in order to transform this gradual notion into a categorical notion of consensus, then this threshold will arguably be lower than 100. (In particular, a society will not be noticeably less stable just because a single citizen does not accept PC.) If so, then the subtuple of belief systems accepting *PC* may be a 'proper' subtuple, i.e. it is different from the tuple itself *even if* there is consensus on PC in that tuple as a whole. As a consequence, there is theoretical possibility that the pluralism value for the subtuple (i.e. *PC*-pluralism) is different from the pluralism value of the tuple itself (i.e. global pluralism). In particular, depending on the relevant pluralism threshold, it is possible that there is global pluralism but not *PC*-pluralism and *vice versa*. We then have to ask ourselves: Which is more important? It seems obvious to me that *PC*-pluralism is the relevant notion here. After all, if there is global pluralism but no pluralism among the belief systems accepting *PC*, then it seems that there is no overlapping consensus in any sense, because there is not a pluralism of doctrines overlapping on *PC*. It is not clear, of course, whether it is a relevant scenario that global pluralism and *PC*-pluralism come apart. But, just in case, I will explicate the notions of a potential global overlapping consensus by requiring the relevant tuples to exhibit a consensus on *PC* and a pluralism of doctrines in their *PC*-subsocieties. (Note that this became necessary only after explicating consensus in a way that allows for these cases.)

Having cleared this point, let's now take the informal definitions and plug into them the following formal ingredients from this chapter:

- Explication 1 of the notion of a belief system as an epistemic state.
- Explications 2 of the notion of a justified epistemic state as a fixpoint of *LocalMRE*. (These are two explications differing in whether the quadratic or linear achievement function is used.)
- Definition 6 of the acceptance rate of a political conception in a tuple of epistemic states. This definition can be used to explicate the (categor-

ical) notion of consensus if one sets a plausible threshold for acceptance rates.

- Definition 7 of the *PC*-subtuple (or *PC*-subsociety) of a tuple of belief systems.

- Definitions 8–10 of different measures for the pluralism of comprehensive doctrines in a subtuple of a tuple of belief systems. Again, these definitions can be used to give different explications of the (categorical) notion of pluralism in the *PC*-subsociety of a tuple of belief systems if one sets plausible thresholds for the different measures.

Using this combination of ingredients results in the following explications:

Let $A = \{a_1, ..., a_n\}$ be a set of agents living together in a society with a shared dialectical structure (S, A). Let $\mathcal{E} = C \times \mathcal{T}$ be the set of possible epistemic states with the set of all minimally consistent positions $C \subset \wp(S)$ (the possible commitments) and the set of all dialectically consistent positions $\mathcal{T} \subset \wp(S)$ (the possible theories). Let $F((S, A), C_0, Alg) \subset \mathcal{E}$ be the set of all possible fixed points (in particular, considering all random choices) of algorithm Alg applied to initial commitments C_0 on dialectical structure (S, A). Let $C_0^{a_i} \in C$ be a_i's initial commitments. Let $t_{AccRate}$ be the lowest plausible threshold for categorical consensus when using the acceptance rate as a measure for gradual consensus, likewise $t_{OptCount}$, t_{SoW} and $t_{Entropy}$ for the corresponding pluralism measures. Let *Pluralism* be a variable that can take three values: *OptCount*, *SoW* and *Entropy*. Let *LocalMRE* be a variable that can take two values: *LocalQuadraticMRE* and *LocalLinearMRE*. Let $PC \in S$ be a political conception of justice. Then:

Explication 3 (Potential global overlapping consensus)**.** There is a *potential global overlapping consensus on PC*

- *in the weak sense* iff there is at least one tuple $\mathfrak{T} \in F((S, A), C_0^{a_1}, LocalMRE) \times \cdots \times F((S, A), C_0^{a_n}, LocalMRE)$ with

$$AccRate(\mathfrak{T}) \geq t_{AccRate}, \text{ and}$$
$$Pluralism(\mathfrak{T}^{PC}) \geq t_{Pluralism}.$$

- *in the strong sense* iff for all tuples $\mathfrak{T} \in F((S,A), C_0^{a_1}, LocalMRE) \times \cdots \times F((S,A), C_0^{a_n}, LocalMRE)$:

$$AccRate(\mathfrak{T}) \geq t_{AccRate}, \text{ and}$$
$$Pluralism(\mathfrak{T}^{PC}) \geq t_{Pluralism}.$$

- *of grade r* iff for a proportion $r \in [0,1]$ of all tuples $\mathfrak{T} \in F((S,A), C_0^{a_1}, LocalMRE) \times \cdots \times F((S,A), C_0^{a_n}, LocalMRE)$, it holds that

$$AccRate(\mathfrak{T}) \geq t_{AccRate}, \text{ and}$$
$$Pluralism(\mathfrak{T}^{PC}) \geq t_{Pluralism}.$$

Explication 4 (Potential local overlapping consensus)**.** There is a *potential local overlapping consensus on PC*

- *in the weak sense* iff there is at least one tuple $\mathfrak{T} \in F((S,A), C_0^{a_1}, LocalMRE) \times \cdots \times F((S,A), C_0^{a_n}, LocalMRE)$ with

$$Pluralism(\mathfrak{T}^{PC}) \geq t_{Pluralism}.$$

- *in the strong sense* iff for all tuples $\mathfrak{T} \in F((S,A), C_0^{a_1}, LocalMRE) \times \cdots \times F((S,A), C_0^{a_n}, LocalMRE)$:

$$Pluralism(\mathfrak{T}^{PC}) \geq t_{Pluralism}.$$

- *of grade r* iff for a proportion $r \in [0,1]$ of all tuples $\mathfrak{T} \in F((S,A), C_0^{a_1}, LocalMRE) \times \cdots \times F((S,A), C_0^{a_n}, LocalMRE)$, it holds that

$$Pluralism(\mathfrak{T}^{PC}) \geq t_{Pluralism}.$$

Phew! These explications are quite a mouthful. I basically just plugged the formal apparatus in the relevant places. Note that I have given these explications presupposing the idealisation that the agents share a dialectical structure, because the explications would have been more complicated otherwise. But the generalisation is straightforward and will be necessary for future studies that are de-idealised in this respect. Also note that these

Real-world entity	Formal representation
Citizen	Initial commitments
Shared dialectical situation	Shared dialectical structure
Society	Shared dialectical structure + Set of initial commitments
Belief system	Epistemic state
State of reflective equilibrium	Achievement
Method of reflective equilibrium	Algorithm
Justified belief systems	Fixpoints of algorithm
Consensus	Acceptance Rate
Pluralism	Option count *or* Strength of the Weak *or* Entropy

Table 3.1: Cheatsheet for translating the informal notions of chapter 2 to the formal notions given in this chapter.

'two' explications actually contain many more, because I used two variables: *LocalMRE* and *Pluralism*. Taking all things together, we have a total of 36 explications of the notion of a potential overlapping consensus! In particular, we have 2 kinds (global and local), 3 senses (weak, strong, graded), 2 algorithms (*LocalQuadraticMRE* and *LocalLinearMRE*) and 3 pluralism measures (option count, strength of the weak, entropy): $2 \times 3 \times 2 \times 3 = 36$.

Let's recap this chapter by discussing a cheatsheet for how the real-world entities are formally represented, see table 3.1. The cheatsheet is based on two important points: First, since I will simulate artificial, not real societies, I have no use for the concept of a belief system that is 'actually held' by a citizen. But the citizens in my simulation study do have initial commitments, thus, every citizen can be represented by their initial commitments. Second, we would theoretically also need the dialectical situation for each citizen, but I here make the idealising assumption that all citizens in a society share the same dialectical situation (represented by a shared dialectical structure). Thus, the dialectical structure is a feature of the society, not of individual citizens. As a consequence, the societies I study are represented by a shared dialectical structure together with the initial commitments of each citizen. In the next chapter I lay out how these artificial societies are constructed.

Next, the belief systems that evolve during equilibration processes, or *hypothetical* equilibration processes as I should stress (see Reconstructionism

from the last chapter), are represented by what are called epistemic states in section 3.2. Each epistemic state is a pair of two positions, the commitments of the agent and the theory of the agent (not actually held, but hypothetically held, of course). The epistemic state's property of being more or less in the state of reflective equilibrium (its degree of equilibrium, see section 2.2.4) is represented by the achievement function. For every epistemic state, the achievement function gives a definite real number between 0 and 1 representing the degree to which this state is in reflective equilibrium. (For doing this, the achievement function needs the initial commitments and the dialectical structure of the agent. That is, this degree is always relative to both of these entities.) The method of reflective equilibrium is represented by the algorithm that optimises the achievement function for every set of initial commitments. More precisely, in my study the equilibration method that is both feasible for any agent and effective at optimising achievement is represented by the local algorithm with neighbourhood depth 1. All epistemic states that can result from the hypothetical application of this algorithm to the initial commitments are propositionally justified for the agent with these initial commitments. In general, there will several such fixpoints. Thus, the set of belief systems that is justified for an agent is represented by the set of fixpoints that can result from applying the algorithm to their initial commitments. This is the basic idea behind the explication of justification given in section 3.3.

As a consequence, given a society of agents, there is a space of justified epistemic states containing all possible combinations of these justified epistemic states. Each such combination is represented by a tuple of epistemic states. For each such tuple, we can ask: Is there a consensus on some political conception? Is there a pluralism of comprehensive doctrines? Both notions are formally represented by functions that take a tuple of epistemic states, or a subtuple thereof, as input and return a real number between 0 and 100. This real number represents the degree of consensus or degree of pluralism. For consensus, this function is the acceptance rate of the tuple. For pluralism, we have three such functions, each focusing on a different aspect of the notion: option count, strength of the weak and entropy. To get explications of the categorical (not gradual) notions of consensus and pluralism, one needs to set thresholds for these functions. I have not done

so thus far and, luckily, we can analyse the results of my simulation study without setting such thresholds. In essence, I will argue that some values of these functions are plausibly above or below the threshold and this will suffice for my purposes. In any case, suppose we have set such thresholds. We have then several precisely specified explications of the concept of overlapping consensus. That is, given the initial commitments and the dialectical structure of the citizens, we have several definite criteria for deciding wether there is an overlapping consensus on a political conception of justice.

Chapter 4

Study design

Now that I explicated the relevant notions, we can have a closer look at how my study is conducted. There is a wide variety of ways to use the model to learn about overlapping consensuses. For example, one could do some empirical research to find out about the common core of the dialectical structures in some existing society and the initial commitments (however interpreted) of its citizens. One could then simulate an RE process for every citizens to see what their justified beliefs are. Do they match the actual beliefs of citizens? Do they form an overlapping consensus? No doubt this would be an interesting endeavour (see also section 4.3). However, the dialectical structure would be big (lots of sentences and arguments) and, depending on the (sample) number of citizens, many simulations would need to be run. Given the current limits of computational power, it would be unfeasible even when using the more frugal local algorithm. More importantly, the results would hold only for the society that was studied. I am here interested in the more general rules for overlapping consensuses.

For these reasons, I adopt a different strategy:

- I consider artificial dialectical structures that are much smaller than real-world examples. Thus, following Hegselmann and Krause (2002, p. 9), I model according to the KISS-principle ("Keep it simple, stupid!"). Nonetheless, I construct these structures in a way that mimics the relevant features of real-world structures, or so I argue in section 4.3. Also, I consider only small sets of agents. The benefit of these restrictions is feasibility, I can run simulations on more structures. However,

this requires idealisations.

- There is a vast space of possible societies conforming to these idealising conditions. It is impossible to simulate all of them. Nonetheless, I wish to get a broad view on this possibility space, for this purpose I will not restrict myself to some small predefined subset. Instead, I randomly sample the possibility space. However, I do not simply sample with a homogeneous probability distribution, as my epistemic interests suggest otherwise.

In what follows I lay out these conditions on dialectical structures and agents and explain how I sample the societies conforming to them.

4.1 Possible Societies

Before I give you a list of abstract conditions, let me explain the main ideas using the example structure in figure 4.1. The structure has 23 sentences plus their negations. The negations are not shown in the graphic. The sentences are ordered along four horizontal lines, starting at the top:

- Line 1, PPS1–PPS4: These are *political particular statements* (short 'PPSs'). These sentences are about particular matters of constitutional essentials. Example: 'The law permitting slavery is unjust'.

- Line 2, PC1–PC3: These are *political conceptions of justice* (short 'PCs'). These sentences are general theories of constitutional essentials. Example: Rawls's lexically ordered principles of justice (this is a rather complex sentence).

- Line 3, CD1–CD4: These are *comprehensive moral doctrines* (short 'CDs'). These sentences are general moral theories that might or might not include constitutional essentials. Example: A virtue ethical theory. Note: If a doctrine does not extend to constitutional essentials, it is called partially comprehensive, otherwise fully comprehensive (following Rawls, PL, p. 13). Arguably, a virtue ethical theory is partially comprehensive (Rogers, 2020).

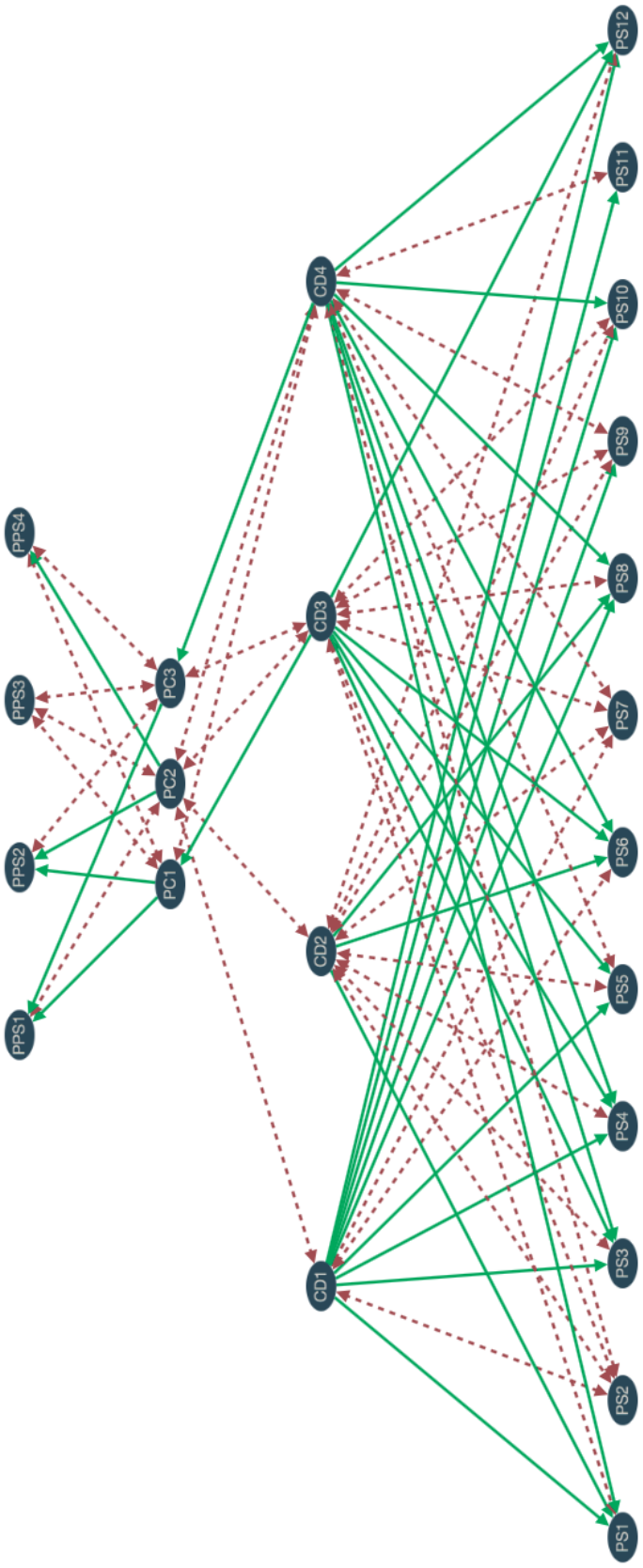

Figure 4.1: Legend: Solid green arrow = implication, dashed red arrow = incompatibility, dotted yellow arrow = joint exhaustiveness.

- Line 4, PS1–PS12: These are *particular statements* (short 'PSs'). These sentences are particular statements about morality, not including constitutional essentials. Example: 'It was wrong of me to lie to my friend'.

All these sentences have in common that they are about morality. They differ along at least two dimensions:

- Generality: Lines 1 and 4 are particular statements, lines 2 and 3 are general statements.
- Subject matter: Lines 1 and 2 are about constitutional essentials. Line 4 is about morality in general without constitutional essentials. Line 3 is also about morality in general but can include constitutional essentials as well.

In figure 4.1 you can also see lots of inferential connections between the sentences. For any pair of sentences, exactly one of three cases obtains:

- A and B are incompatible: A red dashed two-headed arrow (incompatibility is symmetric by contraposition). Formal representation (see sec. 3.1): $\{\{A\}, \neg B\}$ or $\{\{B\}, \neg A\}$. Shorthand: 'incompatibility connection' or 'i-connection'.
- A implies B: A green solid one-headed arrow indicating the direction of implication. Formal representation: $\{\{A\}, B\}$ or $\{\{\neg B\}, \neg A\}$. Shorthand: 'support connection' or 's-connection'. I use 'implication' and 'support' interchangeably.
- There is no connection between the two. In this graphic, there is no arrow for this case, but let's call it the *neutral* connection. It may seem a little odd, but it will soon be very convenient to count this as a connection type. Shorthand: 'neutral connection' or 'n-connection'.

Note that by using this convention we can reference inferential relations to negations of sentences without explicitly mentioning the negations themselves or depicting them in a graphic. Also note that there is one missing though logically possible connection type (between single sentences): $\neg A$ implies B (also symmetric by contraposition, formally: $\{\{\neg A\}, B\}$ or $\{\{\neg B\}, A\}$).

Given the above interpretation of the sentences, it simply makes no sense here.

Note that the arguments in the structure are always arguments with exactly one premise. This is for simplicity. Also, as mentioned in section 3.3, the local variant of the algorithm (with a neighbourhood depth of 1) is only good at reaching fixpoints with high achievement when the structure does not feature arguments with more than one premise. However, remember that the deductive inferential relations between the sentences in a dialectical structure can always be interpreted as presupposing implicit uncontroversial supporting premises (sec. 3.1). Thus, Rawls's model case of an overlapping consensus in section 2.1.1 is representable in structures like these, but only by idealising that the additional premises needed for inferring a political conception from a comprehensive doctrine are uncontroversial.

You will have noticed that the sentences are not arbitrarily connected with each other. More precisely, the arguments in the structure fall into two classes:

- The *head* of the structure: These are the arguments connecting the CDs and PCs. The CDs support the PCs or their negations or are neutral. A CD can only have these three connection types: It cannot happen that a PC implies a CD, because a PC is only about constitutional essentials but a CD is (at least in part if not wholly) about general morality without constitutional essentials. Also it cannot happen that the negation of a CD implies a PC or *vice versa*. (What would that even mean? I cannot think of an interpretation.) Thus, there are only these three connection types between CDs and PCs. However, not any combination of these is possible, as we will see soon in section 4.2.1.

- The *body* of the structure: These are the arguments connecting the general statements (CDs and PCs) with particular statements (PSs and PPSs, respectively) within the relavant subject matter (morality without constitutional essentials and constitutional essentials, respectively). The general statements imply particular statements or their negations or are neutral. This reflects the idea that the general statements are interpreted as theories about their subject matter and as such they give verdicts about particular cases when applied. You might ask:

> Why don't the CDs imply PPSs, if the CDs can be (in part) about constitutional essentials? First note that they sometimes do imply the PPS, but only *indirectly* by implying a political conception. Perhaps there are plausible cases where a CD *directly* implies a PPS, but this would further complicate the structures and make it harder to understand the results presented in the next chapter. Thus, it is another simplifying idealisation that there is no such direct implication.

The choice of the names 'head' and 'body' is overcome but too old to change. Don't think too hard about it. Note that there is no pair of sentences between which the implication/support connection can go both ways. Thus, I will often just talk of the support connection without specifying the direction of implication.

This concludes the main ideas concerning sentences and arguments, i.e. dialectical structures. What about the agents operating on these structures, i.e. what about the citizens? Remember from the cheatsheet in section 3.5 that agents are identified by their initial commitments. Since all agents in a society always operate on a shared structure, their initial commitments are a subset of the sentences in that structure. More precisely, I assume that the initial commitments are a minimally consistent subset of all particular statements (PSs and PPSs taken together) and their negations. Why not also general statements? In effect, this is again an idealising assumption to keep things simple, enabling us to better understand and interpret the results. Also, this idealising assumption fits well with the paradigmatic case of initial commitments being intuitions about particular cases (see, e.g., the introduction to Knight, 2023). (Note, however, that Rawls himself explicitly disagrees about this idealising assumption (1974, p. 8).) Finally, the societies in my simulation study will have a size of 30 agents. That is, for every structure there is a set of 30 initial commitments.

Let's sum up these ideas (and fill in some gaps) in a complete list of conditions for the societies in my study:

1. The shared dialectical structure is such that:

   (a) The sentence pool has four classes of statements:

   - 4 PPSs,

- 1, 2 or 3 PCs,
- 4, 6 or 8 CDs,
- 12 PSs;

plus negations. As a consequence, the overall sentence pool size is 21–27 plus negations. The variability in the numbers of PCs and CDs serves to see how they influence the OC-performance. (The umbrella term 'OC-performance' denotes the probability of the different kinds of overlapping consensuses.)

(b) The arguments connecting these sentences are such that:

- There is always exactly one premise.
- There is at least one consistent and complete position.
- For any pair of CD and PS, the CD either implies PS (support connection) or its negation (incompatibility connection) or there is no argument between them (neutral connection).
- For any pair of PC and PPS, the PC either implies PPS (support connection) or its negation (incompatibility connection) or there is no argument between them (neutral connection).
- For any pair of CD and PC, the CD either implies PC (support connection) or its negation (incompatibility connection) or there is no argument between them (neutral connection).
- CDs are pairwise incompatible.
- PCs are pairwise incompatible.
- For each CD and for each PC, there is at least one complete and consistent position containing it. In this sense, CDs and PCs are never self-contradictory.
- There are no arguments between CDs and PPSs, between PCs and PSs, and between PSs and PPSs.

2. There are 30 agents operating on the shared structure. Each agent's initial commitments is a minimally consistent subset of the union of PSs and PPSs and their negations.

4.2 Sampling the possibility space

As foreshadowed in the beginning of this chapter, these conditions, even though quite restrictive and idealising, leave open a vast space of possible societies. Extremely vast, in fact. There are many more such societies than particles in the known universe. Combinatorial explosion is a real problem. It is not and might never be possible to simulate all possibilities. Thus, we need to sample this possibility space. But how are we to do that? We could just randomly sample this finite possibility space with a homogeneous probability distribution. However, my interests suggest a different way of sampling.

4.2.1 Sampling the heads

As explained in section 2.2.6, I am interested in how the connections between CDs and PCs influence the possibility of an overlapping consensus. Using the terminology introduced above, this means that I am interested in how the *head* of the shared structure influences this possibility. Suppose we have $n_{CD} = 4$ CDs and $n_{PC} = 1$ PC. Consider the following two heads. Head H_1 is such that CD1, CD2 and CD4 support the PC, but CD3 is neutral about it. Head H_2 is such that CD1, CD2 and CD3 are incompatible with the PC, but CD4 supports it. Let's represent these heads somewhat more formally by a *tuple of CD-types*. Basically, one can categorise a CD according to the connections it has to the PCs in the structure. Here, where $n_{PC} = 1$, any CD is of exactly one of the following three CD-types: support, incompatibility and neutrality, depending on whether the CD supports, is incompatible with or neutral about the PC. (Of course, in structures with more than one PC the CD-types are more complex, more on this below.) Every head can be uniquely represented by an n_{CD}-tuple of CD-types. Each position in the tuple corresponds to a CD and the element in that position denotes its type. Head H_1 is represented by (s, s, n, s) and head H_2 by (i, i, i, s).

Now, presumably structures with head (s, s, n, s) and structures with head (i, i, i, s) will allow for an overlapping consensus to a different degree or with a different probability. As you know, I am interested in this influence. But of course, whether or not an overlapping consensus occurs with a certain head

will also heavily depend on the rest of the structure (i.e. the body) and the initial commitments of the agents operating on it. And in order to simulate RE processes so that we can check for pluralism and consensus among the fixpoints, we need both body and initial commitments of the agents. So how do we choose them in a way that let's us compare the OC-performance of different heads?

Of course, the answer is: by simulating not just one society with a certain head, but a bunch of them with random bodies and random initial commitments. In this study, for every head I generate 50 random bodies. For each of the 50 resulting structures I generate 30 random initial commitments. (See section 4.2.2 for more on this.) After running the RE processes in these 50 societies, we can calculate the resulting *average* pluralism and *average* consensus for that head. The influence of bodies and initial commitments will average out, enabling us to compare the performance of different heads.

Multisets of CD-types

Now, we could do that for all possible heads. In our example case above, with $n_{CD} = 4$ and $n_{PC} = 1$, there are $3^4 = 81$ possible heads (for different values of n_{CD} and n_{PC} this number is much bigger, of course). However, this is not necessary. We can save a lot of computational power by using a certain trick. Consider again our example head (s, s, n, s). Suppose we swap the third and fourth connection, resulting in head (s, s, s, n). The only difference between these heads is which CDs have which connections. The total number of occurrences for each connection type is the same (3 x s and 1 x n). Nevertheless, for any two given societies with heads (s, s, n, s) and (s, s, s, n), respectively, the OC-performance will likely differ. This is because their bodies and initial commitments might be, and very likely *will* be, asymmetric between CD3 and CD4. Thus, swapping the connection types of CD3 and CD4 will make a difference, because bodies and initial commitments treat CD3 and CD4 differently. However, since for any head we randomly generate many bodies and initial commitments, their influence is averaged out anyways. As a consequence, we can expect the *average* OC-performance for (s, s, n, s) and (s, s, s, n) to be the same. In other words, even though any two particular societies will likely treat CD3

and CD4 asymmetrically, averaging over 50 random societies *does* treat them symmetrically, giving us similar average results.

This simplifies things significantly. Since I am only interested in the average results for each head and the average results will be the same for both (s, s, n, s) and (s, s, s, n), we can just simulate one of them and ignore the other. In fact, we can ignore *all* heads that are only reordering the same CD-types. All that matters is that they have the same number of occurrences for each CD-type. The class of these heads can be uniquely represented by a so-called *multiset of CD-types*. Here the multiset is {s, s, s, n}. Multisets are like tuples without an ordering. Or like sets with repetitions, hence the name. A head is *instantiating* a multiset iff the head's tuple is an ordering of the multiset's elements. Since we can expect the average OC-performance to be the same for all heads instantiating a particular multiset, we can save computational power by choosing just one of them and simulating 50 random societies with this head. We can then treat the average OC-performance of this particular head as representative of the average OC-performance of *all* the heads instantiating the same multiset.

Let's see how much computational power we saved. Again, for $n_{CD} = 4$ and $n_{PC} = 1$, there are $3^4 = 81$ possible heads. However, we only need to simulate one head per multiset. Following Stanley's notation (1997, p. 15), the number of possible multisets of size n_{CD} from m CD-types is

$$\left(\!\binom{m}{n_{CD}}\!\right) = \binom{m + n_{CD} - 1}{n_{CD}} = \frac{(m + n_{CD} - 1)!}{n_{CD}!\,(m-1)!}. \tag{4.1}$$

In our case, $n_{CD} = 4$ and $m = 3$ gives us 15 multisets. This means we only have to simulate 15 instead of 81 heads to get the same amount of relevant information! You can see that the abstraction of talking about multisets has paid off.

CD-types for 2 and 3 PCs

Let's generalise these considerations to heads with two and three PCs. Consider heads with $n_{CD} = 4$ and $n_{PC} = 2$. Every CD has two connections to the PCs, one for PC1 and the other for PC2. Each connection can be of the three types you already know: support, incompatibility and neutrality. For each

CD, this gives us a total of $3^2 = 9$ combinatorically possible CD-types: ss, si, sn, is, ii, in, ns, ni, nn. The first letter of each CD-type denotes the CD's connection to PC1, the second to PC2. Again, any such head can be uniquely represented by a tuple of these CD-types. However, given the conditions on dialectical structures detailed in section 4.1, it does not make sense to include all these CD-types.

First, if a CD is of CD-type ss, i.e. it implies both PC1 and PC2, then it is self-contradictory, because PCs are mutually incompatible. But CDs must not be self-contradictory. Thus, we can exclude ss, it will always violate the conditions on possible structures. Second, sn and ns are redundant: Suppose a CD is of CD-type sn, i.e. it supports PC1 and is neutral about PC2. Since PC1 and PC2 are incompatible, however, PC1 supports the negation of PC2. Thus, the CD also supports the negation of PC2. (This implication is not explicitly represented in the set of arguments, but what matters is that the negation of PC2 is in the content of the CD.) As a consequence, whenever a CD is of CD-type sn or ns, we can just exchange that CD-type with si or is, respectively, and get an equivalent dialectical structure. In particular, this means that all sentences have the same content as before and RE processes will run just the same. Thus, we can reduce complexity and save computational power by excluding sn and ns as CD-types. We are left with six CD-types: si, is, ii, in, ni, nn. (I am indebted to Gregor Betz for this way of simplifying the study design.)

CD-types for $n_{PC} = 3$ work just the same. Here for each CD we have a total of $3^3 = 27$ combinatorically possible connections to the PCs. Again, we can sort out a lot of CD-types following the two above considerations: All CD-types with more than one support connection have to go, otherwise the CD will be self-contradictory. All CD-types with a support connection and at least one neutral connection have to go as well, because they are redundant. This leaves us with eleven CD-types: sii, isi, iis, iii, iin, ini, inn, nii, nin, nni, nnn. For example, the head of the structure in figure 4.1 can be described by the tuple of CD-types (nin, nin, sii, iis).

Again, we can do the trick of focusing on multisets of CD-types instead of tuples of CD-types. Using equation 4.1, we get table 4.1 displaying the total number of possible multisets for each combination of n_{CD} and n_{PC}. Note that for $n_{CD} = 8$ and $n_{PC} = 3$, the number of combinatorically possible

	1 PC	2 PCs	3 PCs
4 CDs	15	126	1.001
6 CDs	28	462	8.008
8 CDs	45	1.287	43.758
		Total	54.730

Table 4.1: Number of combinatorically possible multisets of CD-types.

heads is $3^{8\cdot3} = 282.429.536.481$. We boiled that down to 43.758 multisets, each represented by only one simulated head. We have tamed the beast of combinatorial explosion. Or have we? If we pick one head for each of the 54.730 multisets, generate 50 random bodies and for each resulting structure 30 initial commitments, this results in a total of 82.095.000 agents to be simulated. That's still a little too much. Thus, we will randomly sample these multisets with homogeneous probability distribution, but we will do so separately for each combination of n_{CD} and n_{PC}. Table 4.2 displays the sample sizes. This gives us a total of $3.748 \times 50 \times 30 = 5.622.000$ agents that are

	1 PC	2 PCs	3 PCs
4 CDs	15	126	768
6 CDs	28	462	768
8 CDs	45	768	768
		Total	3.748

Table 4.2: Sizes of the samples of combinatorically possible multisets. The multisets were sampled with homogeneous probability distribution.

simulated. For each of these agents, the local algorithm will be run once with the quadratic achievement function and once with the linear achievement function. This gives a total of 11.244.000 simulated RE processes. Table 4.3 displays an overview of the numbers.

Multisets	3.748
Heads	3.748
Bodies per head	50
Structures	187.400
Agents per society	30
Agents	5.622.000
RE processes	11.244.000

Table 4.3: The simulation study in numbers.

The attentive reader will have noticed that I said that for each agent the algorithm will be run *once* for both quadratic and linear achievement function. But wait! Didn't we say in the previous chapters that for every agent there will likely be several possible fixpoints and not just one? Yes, we did say that. The combinations of these sets of fixpoints form the space of justified belief systems in a society. And there will be a potential global or local overlapping consensus in the different senses depending on how many tuples of belief systems in this space exhibit certain properties (i.e. pluralism and consensus). Thus, one might think, in order to judge how many such tuples there are we would need to find not one only such fixpoint for each agent, but all of them. This is, in principle, true. However, it is extremely costly in terms of computational power to find all fixpoints for each agent. For this reason, the present study will only find one such fixpoint for each agent. As a consequence, for each society we will have only one tuple of justified belief systems and not the whole space. Luckily, as you will see in section 5.2, we will nonetheless be able to draw interesting conclusions from the resulting data. But still, it is important to keep in mind right from the start that by running the algorithm only once (though with both quadratic and linear achievement function) we effectively draw one random tuple from the space of justified belief systems, though not necessarily with a homogeneous probability distribution. I will return to this point when interpreting the results of the study.

4.2.2 Sampling bodies and initial commitments

I have not yet detailed how bodies and initial commitments are sampled. The most straightforward way of doing this is to sample the possibility space of the bodies and the initial commitments with homogeneous probability distribution:

- Construct a random body:
  - For any pair of PC and PPS, realise either the support, incompatibility or neutral connection with equal probability and add it to the set of arguments.

  - For any pair of CD and PS, realise either the support, incompatibility or neutral connection with equal probability and add it to the set of arguments.

  Repeat 50 times to get 50 bodies.

- Construct a random set of initial commitments: For every PS and PPS, decide at random whether the initial commitments contain it, its negation or neither. Repeat 30 times to get 30 sets of initial commitments.

For the initial commitments we will do just that. For the bodies, however, a slightly different approach seems more appropriate.

Note that any CD will be expected to imply a third of the PS, be incompatible with another third and neutral about another third, the same holds for the PCs and PPSs. However, since the CDs are supposed to be *comprehensive* moral doctrines, it seems that being neutral about a third of the PSs is a bit too much neutrality. For this reason, I will tweak the probabilities such that there is a 45% chance for s, 45% for i and 10% for n. Thus, the expected neutrality for any CD is 10%. We could do the same for the PCs and PPSs. However, since it is only 4 PPSs to begin with, a PC that is neutral about only one PPS (the minimal amount) will have 25% neutrality. This seems a bit much already. A political conception of justice that leaves open one fourth of the critical questions on constitutional essentials seems too non-committal. So let's idealise a bit and assume that PCs cannot be neutral about the PPSs, i.e. the probabilities are 50% for s and 50% for i. In a later study design with more PPSs, we may drop this idealisation.

As a consequence of tweaking these probabilities, we will sample the possibility space with an *inhomogeneous* probability distribution (indeed completely excluding some of the possibilities), but there are good reasons for doing so.

4.3 What about the real world?

Given that I randomly generate societies from scratch, you might wonder: What about the real world? How do the simulation results for artificial toy societies give us information about what we are actually interested in? First,

I want to stress that I do not generate completely random societies, but I give certain boundary conditions (section 4.1). These boundary conditions are designed such that they mirror features of real-world societies. In particular, I am thinking of the heads that are structurally similar to (parts of) the dialectical situations of actual citizens in actual societies: Most if not all societies do have some kind of public debate. This public debate will to some extent be about different worldviews (the comprehensive doctrines) and different views on constitutional essentials (the political conceptions of justice) and the connections between these two 'levels'. Of course, dialectical situations in the real world are much messier than the clear-cut structures I generate for the study. In particular, they will include more than just doctrines, conceptions and a bunch of particular statements related to them. As I have stressed many times already, it remains to be seen whether the results of the study are robust when the design is de-idealised. But the design presented in this chapter seems like a plausible starting point. In section 6.3, I discuss next steps for modifying and extending the model.

Additionally, the (modified and extended) model can be applied to empirical data. If structures and initial commitments are randomly generated, as I do right now, then the findings hold for the possibility space as a whole. This by itself is, of course, a valuable insight. It helps us understand how MRE generally works given certain boundary conditions. As I explained in the introduction: If there is no defeating evidence, we can infer that MRE works similarly in real cases. This is not unlike the statistical inference from studies about drug efficacy to what can be expected in individual cases. Nonetheless, it is always possible that the more realistic subset of the possibility space as a whole shows a somewhat different behaviour, just like a certain class of individuals might react differently to a drug than the population as a whole. Thus, it will be worthwhile to conduct empirical studies about the following: First, what kind of structures underlie the dialectical situation in particular real societies? Second, what kind of initial commitments do real citizens have in these societies? We can then pair these empirical boundary conditions with simulations of MRE to see whether and how an overlapping consensus is possible. In principle, this can go both ways, we might find that a consensus is easier to achieve than in possibility space as a whole, or harder.

In this context I also want to return to a point that regularly came up in chapter 2: The point that I am adopting a purely structural perspective and, as a consequence, can be non-committal with respect to many hotly debated philosophical questions. I hope that, given the last two chapters, it is now much clearer what I mean by 'purely structural perspective'. I simply don't interpret the sentences in the structures and initial commitments. I am not concerned with their content, only with the inferential relations between them. In this respect, the present research is similar to investigations into the logics of a subject matter. In a sense, this is not surprising, since I am doing formal epistemology.

The consequence, i.e. that I can be non-committal with respect to many controversial philosophical questions, is clearly a strong suit of this approach. It does not matter what you think a 'comprehensive doctrine' or a 'political conception of justice' is, or what divides the purely political from other moral questions, or what you think the appropriate starting point for equilibration processes is, etc. At the end of the day, if you share the general picture of equilibrationist justification, then there will be some initial commitments and there will be a dialectical situation, including whatever you think a comprehensive doctrine or a political conception of justice is, and thus the respective results will be relevant for you. (Of course, as soon as one applies the model to the real world, as I sketched in the last paragraph, one needs to take a stance on these matters.)

In fact, I think that the results might also be relevant for areas other that political liberalism, e.g. the discussion about mid-level moral principles (Bayles, 1986; Espinoza and Peterson, 2012). In a sense, political conceptions of justice are mid-level moral principles: They are not comprehensive moral theories, but also not particular statements. And, since I adopt a purely structural perspective, who says we *must* interpret the respective sentences as political conceptions? Instead, they can also represent other mid-level moral principles. As a consequence, the results of the present study might also give an answer to the following question: What kind of inferential connections between competing general theories on the one hand and competing mid-level principles on the other hand make a consensus on a mid-level principle possible despite a pluralism concerning the general theories? This is and interesting an relevant question. Indeed, I just formulated this question

without using the term 'moral' at all. In interpreting the study results, we might not even be bound to moral philosophy. If you think that MRE gives a plausible account of scientific or even general epistemic justification (as, e.g., Elgin (2005) does), then we can apply the results to other fields like philosophy of science.

Let's recap this chapter. In the previous chapter I have presented a formal model of the method of reflective equilibrium and used the model to explicate the notion of justification, alongside with explications of the different notions of overlapping consensuses. In this chapter, I have detailed how I wish to address the research question and hypotheses by simulating equilibration processes in artificial societies.

First, I characterised these artificial societies by giving a list of necessary conditions for them. These are the most important points:

- Small societies, i.e. 30 agents, each represented by their initial commitments. This saves computational power and allows to simulate more societies.

- Shared dialectical structures. This is an idealisation resulting from my interest in how the *common core* of citizens' dialectical situations influence the possibility of an overlapping consensus.

- Small dialectical structures focusing on the inferential relations between comprehensive doctrines and political conceptions of justice. The set of these inferential relations I have called the structure's *head*.

Thus, every society in my study can be represented by a structure accompanied by a set of initial commitments for the citizens.

Second, I described how I wish to sample the possibility space of artificial societies. The focus of this sampling procedure is on the heads, because it is their influence I am interested in. In particular, for every head I generate 50 random bodies (comprising the rest of the inferential relations). For each of the 50 resulting structures I generate 30 initial commitments. Since I am only interested in average values for the heads (e.g. average consensus, average pluralism), I can save a lot of computational power by considering the equivalence class of heads with the same number of CD-types, i.e. heads instantiating the same multiset of CD-types. For each such multiset, one

head is chosen and its average values are representative of the whole class. Finally, since the number of multisets is still too great to simulate all of them, for each combination of n_{CD} and n_{PC} I randomly sample all possible multisets with homogeneous probability distribution. Each of the resulting societies is simulated once with *LocalQuadraticMRE* and once with *LocalLinearMRE*. As a consequence, for each society we will randomly draw one tuple from the space of justified belief systems given by *LocalQuadraticMRE* and another tuple from the one given by *LocalLinearMRE*.

Now, let's have a look at the results.

Chapter 5

Simulations

In this chapter I will present the results of the simulation study (section 5.1) and discuss what they tell us about the research hypotheses (section 5.2). I should note right from start, however, that this chapter will be concerned exclusively with the technical results of the study. I will take up their *philosophical* interpretation in section 6.2.

Throughout this chapter, we will frequently need the different definitions and explications from chapter 3. I restated the most important ones collectively in appendix C for quick access.

5.1 Results

A central goal of this thesis is to test hypotheses L and G (see section 2.2.6). These are about a particular (though arbitrary) PC. In most structures of this study there is more than one PC. Thus, in order to test one of the hypotheses, we must focus on a fixed PC and check whether the hypothesis holds for that PC. If the sampling is good, then the average results should be similar for any fixed PC (and indeed they are, see robustness result D in the appendix). Thus, we can choose how we like. Since PC1 occurs in all structures (while PC2 and PC3 only occur in structures with two and three PCs, respectively), we will get the most information by considering PC1 as the fixed PC.

The two main goals of this first section are to

1. understand how the inferential connections between CDs and a given PC influence consensus and PC-pluralism, and

2. check whether the results of the study are by and large plausible, i.e. they can be explained and understood without making implausible assumptions.

It will turn out that, with some exceptions, the results are by and large plausible. In the following section, I investigate what they tell us about the research hypotheses.

5.1.1 Ternary heatmaps

In order to test hypotheses L1–3 and G1–3, we need to categorise the structures in my study according to the connections that the CDs have to PC1. These connections are part of a structure's head, thus, we will categorise the heads. Remember that the heads were identified by multisets of CD-types. With respect to PC1, the CD-types can be of three categories: i, n, and s, depending on whether a CD of this type is incompatible with, neutral about or supportive of PC1. For example, for $n_{PC} = 2$ there are six CD-types: si, is, ii, in, ni, nn. The types is, ii, in fall into the incompatible category, because CDs of this type are incompatible with PC1. Types ni and nn fall into the neutral category and si into the support category. By counting how many of the CD-types in a given structure fall into the incompatible, neutral and support categories, we categorise the structures according to the connections between the CDs and PC1 in particular, instead of the connections between the CDs and the PCs in general (as we did for the heads themselves).

Since there are only three such categories for the CD-types and the total number of CD-types always equals n_{CD}, we can use so-called *ternary heatmaps* for visualising the properties of these heads. Ternary heatmaps lump together societies that have heads with the same connections to PC1 into *bins*. The 'heat' of each bin corresponds to some average value for the societies in that bin. Have a look at figure 5.2, displaying the arithmetic mean acceptance rates for PC1. I separate the data for different achievement functions (quadratic and linear) as well as different values for n_{CD} (4, 6 and 8) and n_{PC} (1, 2 and 3). This gives us $2 \times 3 \times 3 = 18$ ternary heatmaps. Every ternary heatmap is an equilateral triangle filled with coloured hexagons. Each such hexagon corresponds to specific numbers of s-, i- and n-connections to PC1 and serves as a bin for all heads with these connections to PC1. The colour of

a hexagon (i.e. its 'heat') indicates the average acceptance rate of PC1 for all heads with that combination. (Note that some hexagons are empty, because the heads are randomly sampled (as described in the last chapter) and there is no guarantee that for each hexagon there is at least one sampled head.)

As an example, consider the ternary plot in the first row and first column in figure 5.2. This heatmap contains all societies with $n_{CD} = 4$, $n_{PC} = 1$ that were simulated using the quadratic achievement function. The three corners of the triangle correspond to the three connection types 'incompatible' (lower left), 'neutral' (top) and 'support' (lower right). The hexagon in the lower left corner contains all heads with only i-connections to PC1. Let's call this hexagon the 4i0n0s-hexagon, because it contains all heads with 4 i-, 0 n- and 0 s-connections to PC1. The hexagon in the top corner (0i4n0s) contains all the heads with only n-connections. The hexagon in the lower right corner (0i0n4s) contains all the heads with only s-connections. I've added text labels to the corners to help remember this convention.

Now, if you start in the i-corner (lower left) and go one hexagon closer to the n-corner (top), then that hexagon will contain all heads with three i-connections to PC1 and one n-connection to PC1. In a sense, by going that step towards the n-corner you have exchanged one of the i-connections by an n-connection. The resulting hexagon is denoted by 3i1n0s. If, instead, you go one step towards the s-corner (lower right), then that hexagon will contain all heads with three i-connections and one s-connection (3i0n1s). By making such steps through the ternary plot, you can reach hexagons with all possible numbers of i-, n- and s-connections to PC1. As a general rule of thumb, the closer a hexagon is to the n-corner, the more n-connections will the heads in that hexagon have, likewise for the other connection types.

Lastly, let me introduce the concept of *isolines*. In the sense that I am using the concept, isolines are sets of hexagons with the same number of connections of some type. For example, the hexagons on the left side of the triangle all have 0 s-connections. Thus, they form an *s-isoline*. Let's call this s-isoline the 0s-isoline. The next parallel line of hexagons that is one step closer to the support corner is called the 1s-isoline. By comparing hexagons on an isoline, we can isolate the influence of two connection types for a fixed number of the third. (Note that my use of the concept of isolines deviates somewhat from the ordinary use. Typically, isolines denote lines

with a constant *target* value. In our case of ternary heatmaps, the ordinary use would be to denote lines of constant heat. However, I use it to navigate the ternary plots, i.e. isolines always denote the same hexagons in the plot, no matter their heat.)

In figure 5.1, I give an overview of how the study is structured, including the hexagons.

Now, let's discuss these ternary heatmaps for different arithmetic mean values. Figure 5.2 shows the average acceptance rates, figures 5.4–5.6 show the average PC1-pluralism for different pluralism measures.

5.1.2 Consensus

Let's start with the arithmetic mean acceptance rates. I first describe the ternaries, then attempt an explanation of the findings.

Description of the results

First and foremost, all ternary plots look strikingly similar. This indicates that the model variant (linear vs. quadratic) as well as the total numbers of comprehensive doctrines (4, 6, 8) and political conceptions (1, 2, 3) do not make a huge difference regarding the influence of PC1 connection types on acceptance rates. Here are some similarities shared by all acceptance rate ternaries:

SA1 The biggest factor by far seems to be the number of s-connections to PC1. The closer one gets to the s-corner (no matter from which direction), the higher the average acceptance rate of the hexagon. Interestingly, there are no huge differences on the s-isolines. That is, given a particular number of s-connections to PC1, the numbers of i- vs. n-connections for the remaining CDs do not make a big difference.

SA2 However, it does make somewhat of a difference. In particular, n-connections to PC1 seem to be better for acceptance of PC1 than i-connections. This influence is particularly visible on the 0s- and 1s-isolines.

Even though the ternaries look mostly similar, there are some differences:

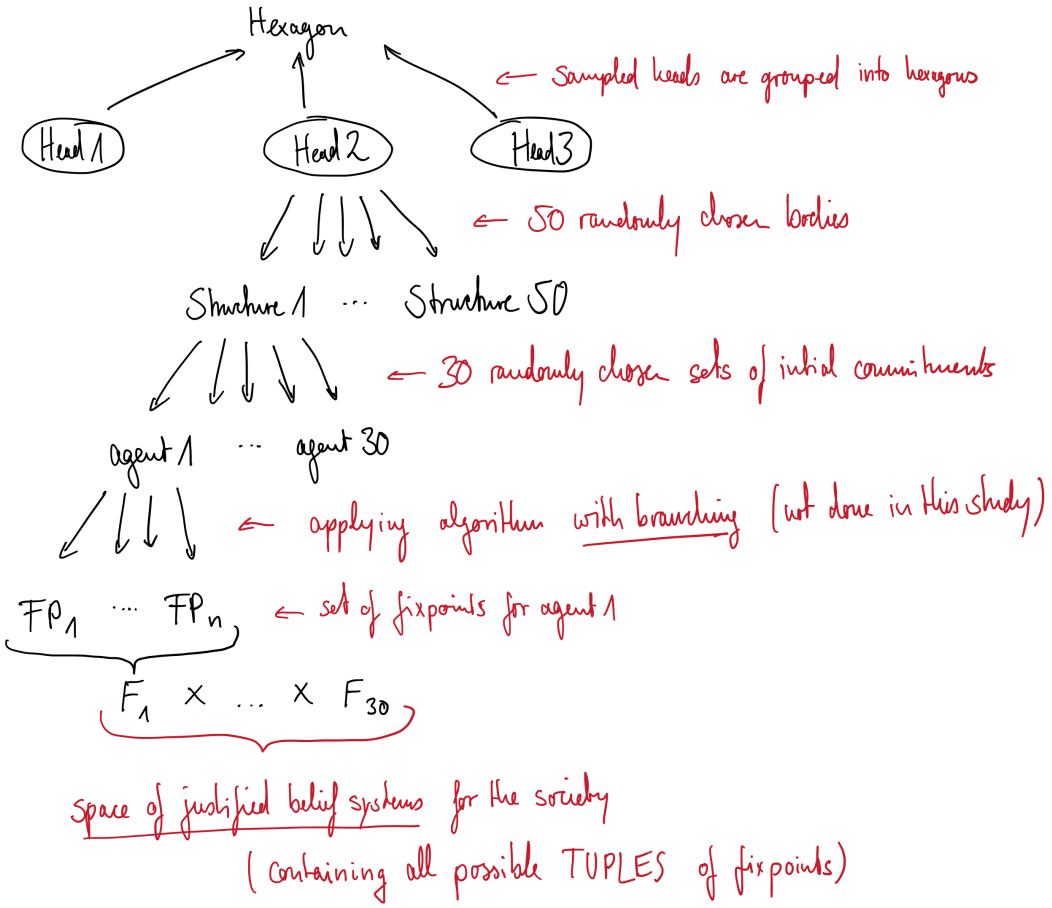

Figure 5.1: Overview of the structure of the study. Heads fall into hexagons depending on the connections of the CDs to PC1. Each head consists of 50 societies. Each society consists of a shared dialectical structure and 30 sets of initial commitments, i.e. 30 agents. For each agent, both the linear and quadratic version of LocalMRE is applied. If it were applied *with branching*, then all possible fixpoints for each agent would be calculated. The Cartesian product of these 30 sets of fixpoints (one set for each agent) is the space of justified belief systems of that society. However, I apply LocalMRE *without branching*. Thus, only one fixpoint per agent is randomly chosen (with unclear probability distribution). As a consequence, only one tuple of fixpoints from the space of justified belief systems is randomly chosen (with unclear probability distribution). For this tuple, we can calculate its acceptance rate, entropy, strength of the weak, and option count. In order to get, e.g., the arithmetic mean acceptance rate of some hexagon (its 'heat'), we average over all 50 tuples per head (one tuple per society and 50 societies per head) and over all heads per hexagon. Every society contributes precisely twice in the 18 ternary heatmaps: Once in a hexagon of a ternary heatmap where its tuple of fixpoints was calculated using the *quadratic* version of LocalMRE, and once in the corresponding hexagon of the accompanying ternary heatmap to the right (same n_{CD} and n_{PC}) where its tuple of fixpoints was calculated using the *linear* version of LocalMRE.

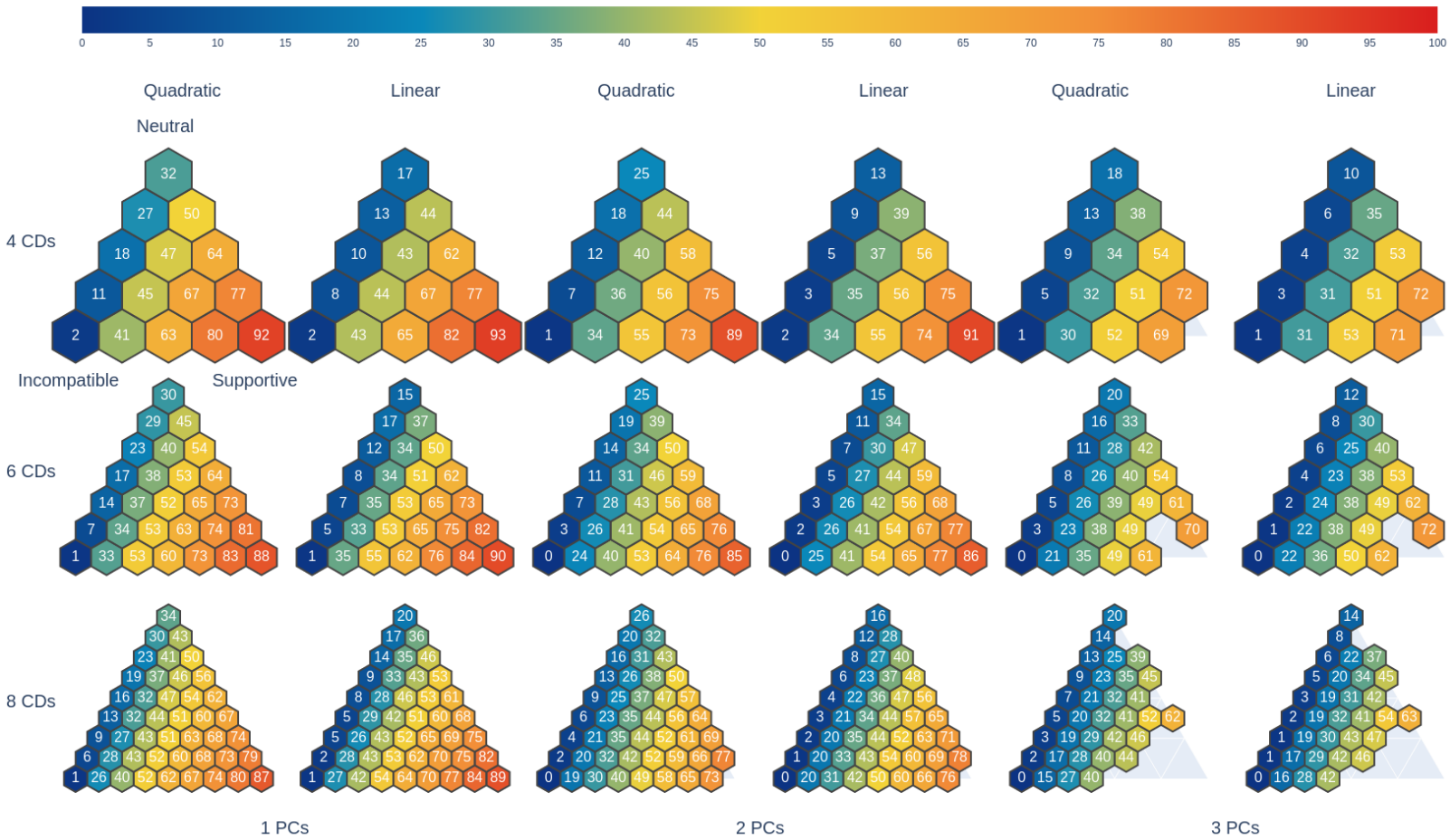

Figure 5.2: These ternary plots display the arithmetic averages of the *acceptance rates*. The data is split up according to model variant, n_{CD} and n_{PC}.

DA1 When comparing specific hexagons between ternaries with the same n_{CD} and the same model variant, e.g. the 2i1n3s hexagon for $n_{CD} = 6$ and linear model variant, then the acceptance rate decreases as n_{PC} increases. That is, more rival political conceptions make acceptance of PC1 less likely.

DA2 The hexagons around the n-corner have a higher acceptance rate for the quadratic model when compared to the linear model. That is, the positive influence of neutral connections (compared to incompatibility connections) on acceptance of PC1 is stronger in the quadratic model.

Before discussing potential explanations for these findings, let's analyse the acceptance rates further. There are only three mechanisms M1–3 that can lead to acceptance of PC1:

M1 Some CD is accepted in the theory in order to account for the agent's commitments. The CD supports PC1. In order to maximise account when adjusting the commitments, PC1 is added to the commitments.

M2 Some CD is accepted in the theory in order to account for the agent's commitments. The CD is neutral about PC1. Nonetheless, since PC1 accounts well for the commitments in the purely political part of the structure (i.e. the PPS-commitments), PC1 is added to the theory as a principle in order to increase account. As a consequence, it is added to the commitments as well.

M3 No CD is accepted in the theory, because none of them account well enough for the commitments. Nonetheless, since PC1 accounts well for the commitments in the purely political part of the structure, PC1 is added to the theory as a principle in order to increase account. As a consequence, it is added to the commitments as well.

The data shows that this list of mechanisms is indeed exhaustive, i.e. every process leading to acceptance of PC1 in the fixed point commitments falls in one of these three categories. Let's call them M1-, M2- and M3-acceptance, respectively. (In appendix E, I explain why there is no other way for PC1 to end up in the fixpoint commitments.) Taking the two ternaries (quadratic and linear model) for $n_{CD} = 6$, $n_{PC} = 2$ as an example, I have split up the

total acceptance rate for PC1 into the respective contributions of the three mechanisms. The results are displayed in figure 5.3. It is immediately clear that the main mechanism contributing to the acceptance of PC1 is M1. One can see that the percentage of RE processes leading to M1-acceptance rises with the number of s-connections. There is no difference on the s-isolines. Interestingly, the difference between the 0s- and 1s-isolines is quite large. M1-acceptance rises from 0% to 23% of the processes, just because one out of six (!) CDs now has an s-connection to PC1. When further moving to the s-corner, M1-acceptance keeps increasing, but not as much. Regarding M2-acceptance, there is a huge difference between the model variants. It almost never occurs in the linear model, but in the quadratic model it does occur relatively often in the area around the n-corner, with up to 20% of processes leading to M2-acceptance in the n-corner itself. Regarding M3-acceptance, there is again a big difference between the model variants. M3-acceptance almost never occurs in the quadratic model, save for a few processes on and around the 0i-isoline. In the linear model, M3-acceptance occurs around the n-corner with up to 10% of the processes in the n-corner itself.

Potential Explanation

In what follows, I attempt to explain these findings.

SA1 The big hotspot around the s-corner is easily explained by the fact that M1 is the main mechanism leading to acceptance of PC1. Thus, s-connections are the best predictor for a high acceptance rate.

SA2 The fact that n-connections are better for acceptance than i-connections is explained by the fact that both M2 and M3 have hotspots around the n-corner. For the quadratic model it's mainly M2, for the linear model it's mainly M3.

DA1 The fact that the acceptance rates for specific hexagons (with constant n_{CD} and model variant) decrease as n_{PC} increases is explained by the following:

- For $n_{PC} > 1$, M2 and M3 do not in principle make a difference between PCs. That is, given that a neutral CD (M2) or no CD (M3)

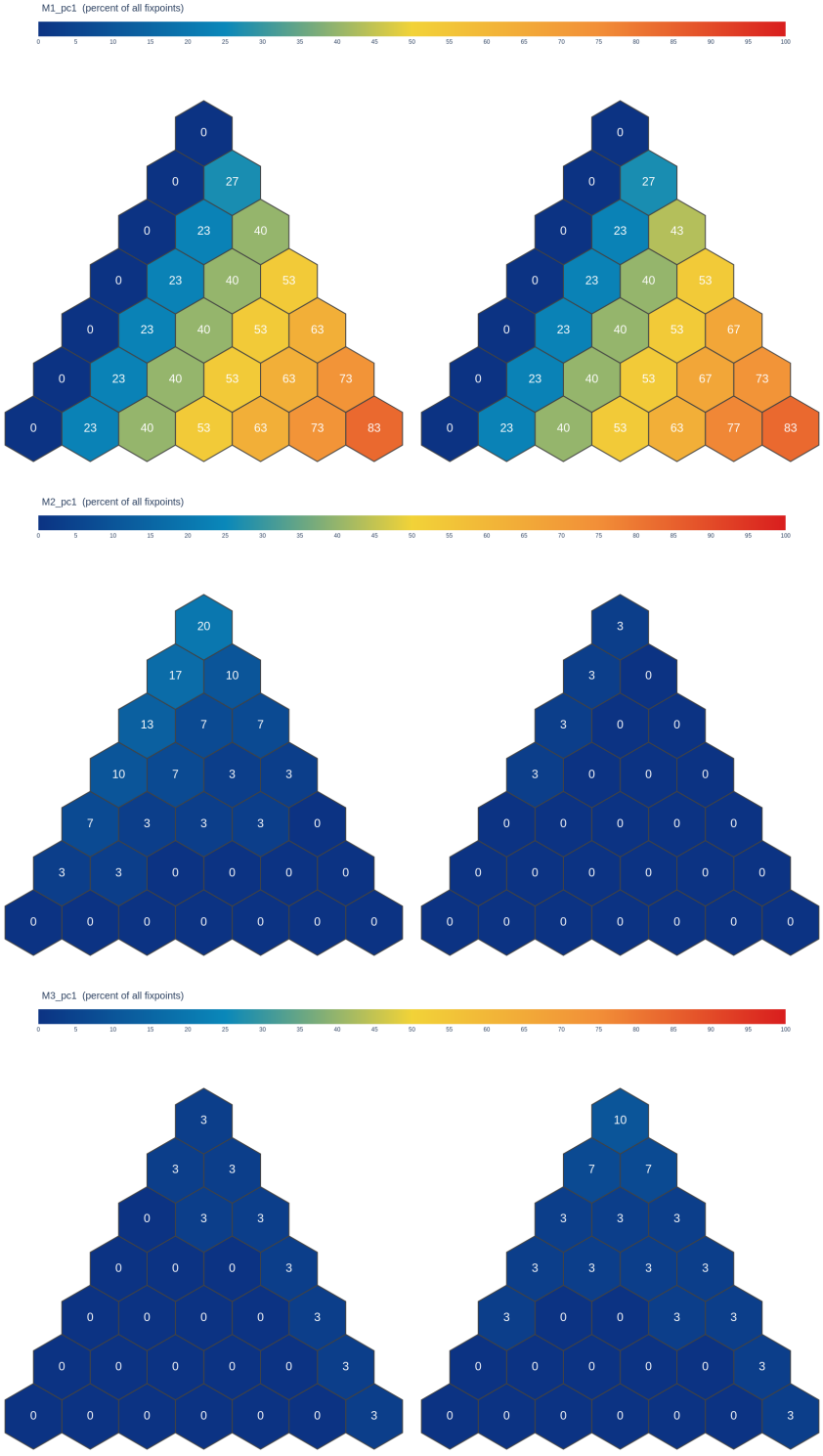

Figure 5.3: These three pairs of ternaries show the arithmetic averages for *M1-, M2- and M3-acceptance* (from top to bottom), in the societies with $n_{CD} = 6$, $n_{PC} = 2$. The left of each pair shows the averages for the quadratic model, the right shows them for the linear model. Note that these numbers are just the acceptance rates split up. Thus, they don't sum to 100, but to the average overall acceptance rate in the respective hexagon.

fits well with the agent's PS-commitments, it can't be taken for granted that PC1 fits best with the agent's PPS-commitments. In particular, as n_{PC} increases, there are more potentially better alternatives to account for the PPS-commitments of the agent. Thus, the probability that PC1 is chosen to account for them decreases. As a consequence, M2- and M3-acceptance can be expected to decrease with n_{PC}.

- M1, however, should be independent of n_{PC}. That is, since M1 overwhelms M2 and M3 as we get closer to the s-corner, the decrease should become smaller as we get closer to that corner. And indeed, this seems to be the case.

DA2 The hexagons around the n-corner have a higher acceptance rate for the quadratic model when compared to the linear model. This follows directly from the fact that in the quadratic model M2 is responsible for acceptance in the n-corner (with up to 20% for $n_{PC} = 2$, see fig. 5.3) while in the linear model M3 is responsible for acceptance in the n-corner, but with a significantly lower contribution (up to 10% for $n_{PC} = 2$).

This concludes my discussion of the acceptance rates. Of course, for a deeper understanding we would have to have a closer look at why M1–3 are distributed over the ternaries in the way that they are. Nonetheless, I think that the results for acceptance rates are by and large plausible. At least, there is no weird or implausible feature that immediately catches the eye.

5.1.3 Pluralism

Let's turn to the pluralism ternaries. Figures 5.4–5.6 show the arithmetic mean pluralism of CDs in the PC1-subsociety. In figure 5.4, pluralism was measured as entropy, in figure 5.5 as strength of the weak, and in figure 5.6 as option count (see section 3.4 for the definitions).

Description of the results

Again, the ternaries look rather similar, however, there are some significant differences. Let's start with the similarities:

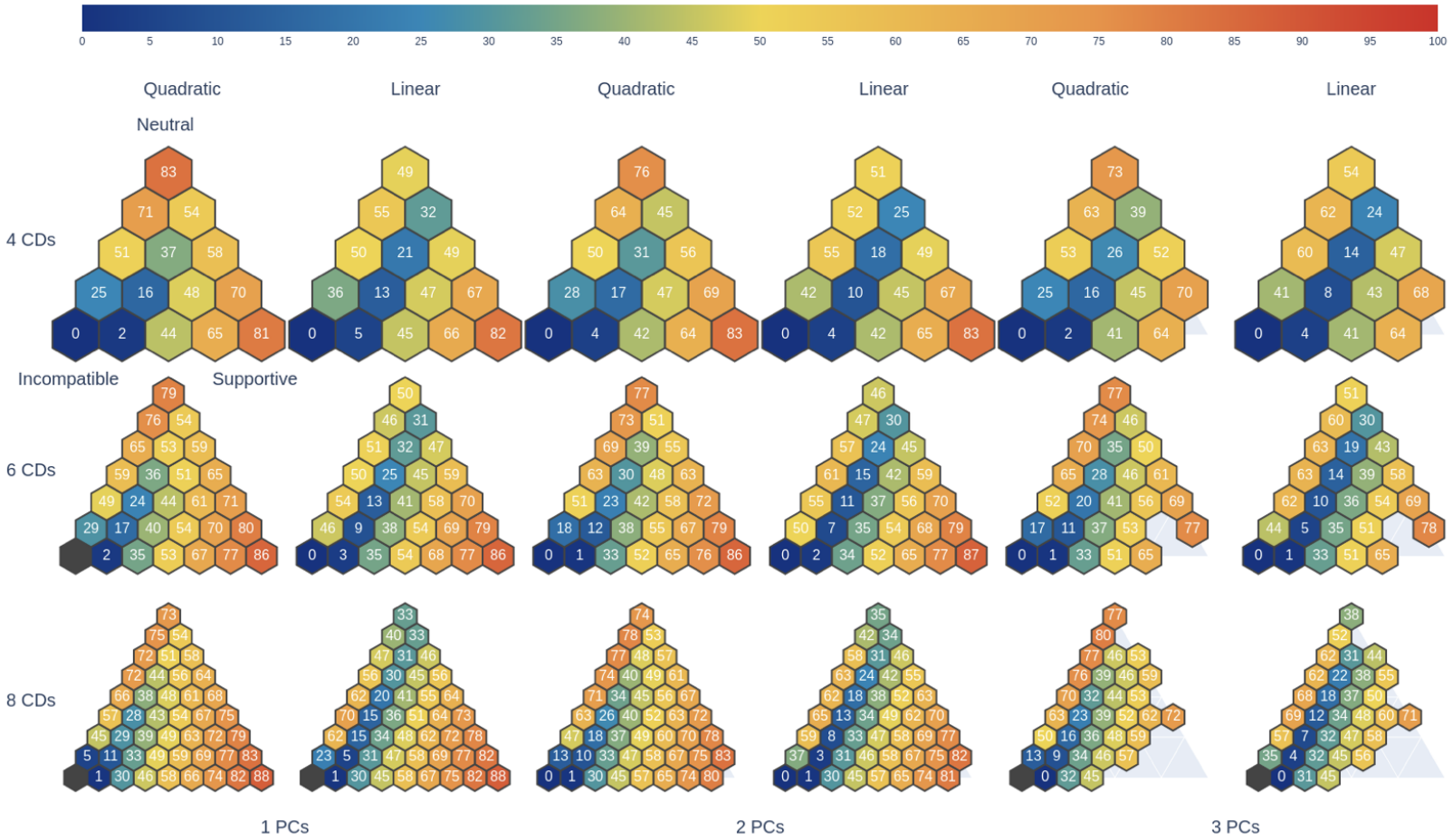

Figure 5.4: These ternary plots display the arithmetic averages of the *entropy* in the PC1-subsociety of the drawn tuples. The data is split up according to model variant, n_{CD} and n_{PC}.

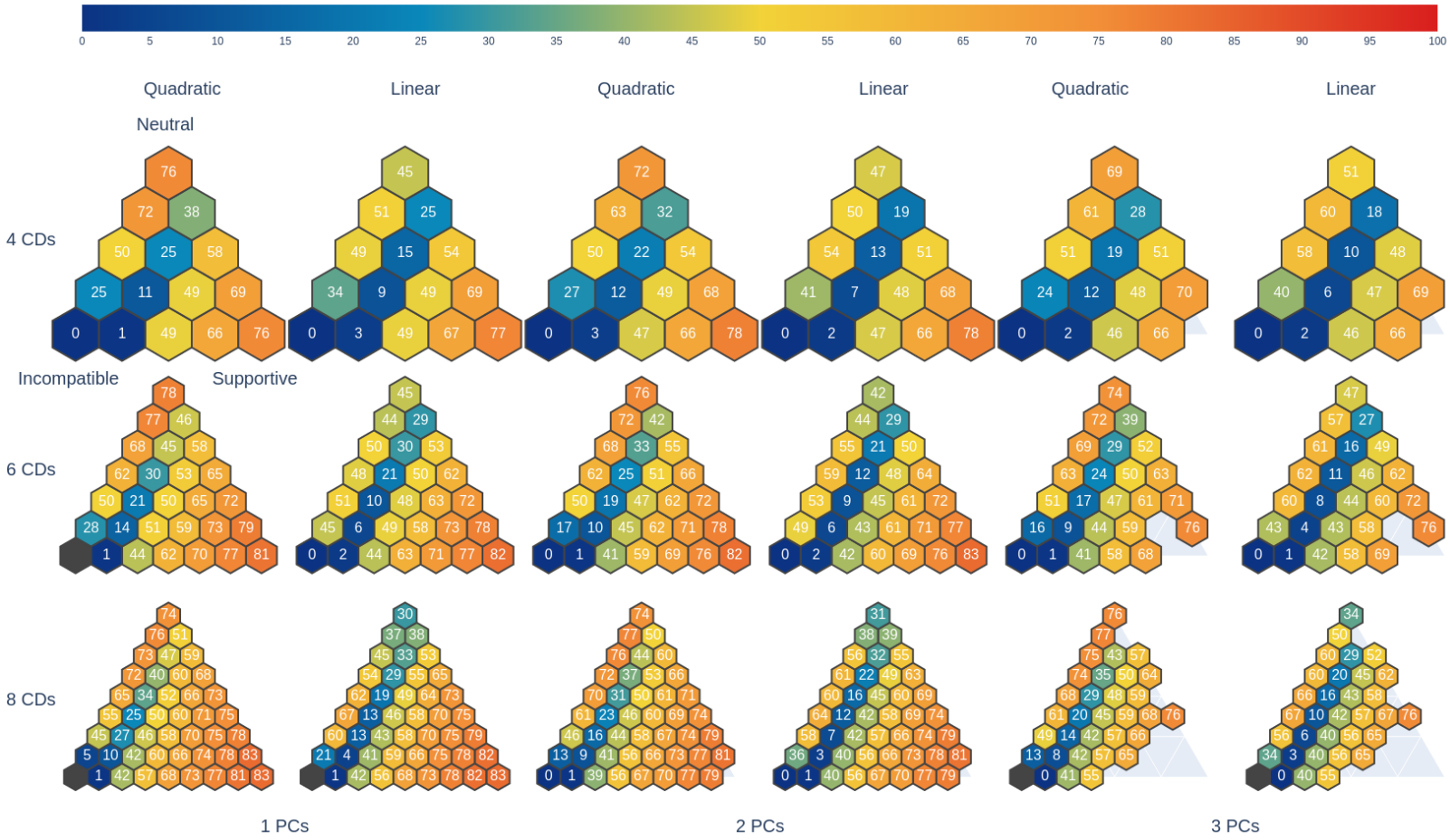

Figure 5.5: These ternary plots display the arithmetic averages of *strength of the weak* in the PC1-subsociety of the drawn tuples. The data is split up according to model variant, n_{CD} and n_{PC}.

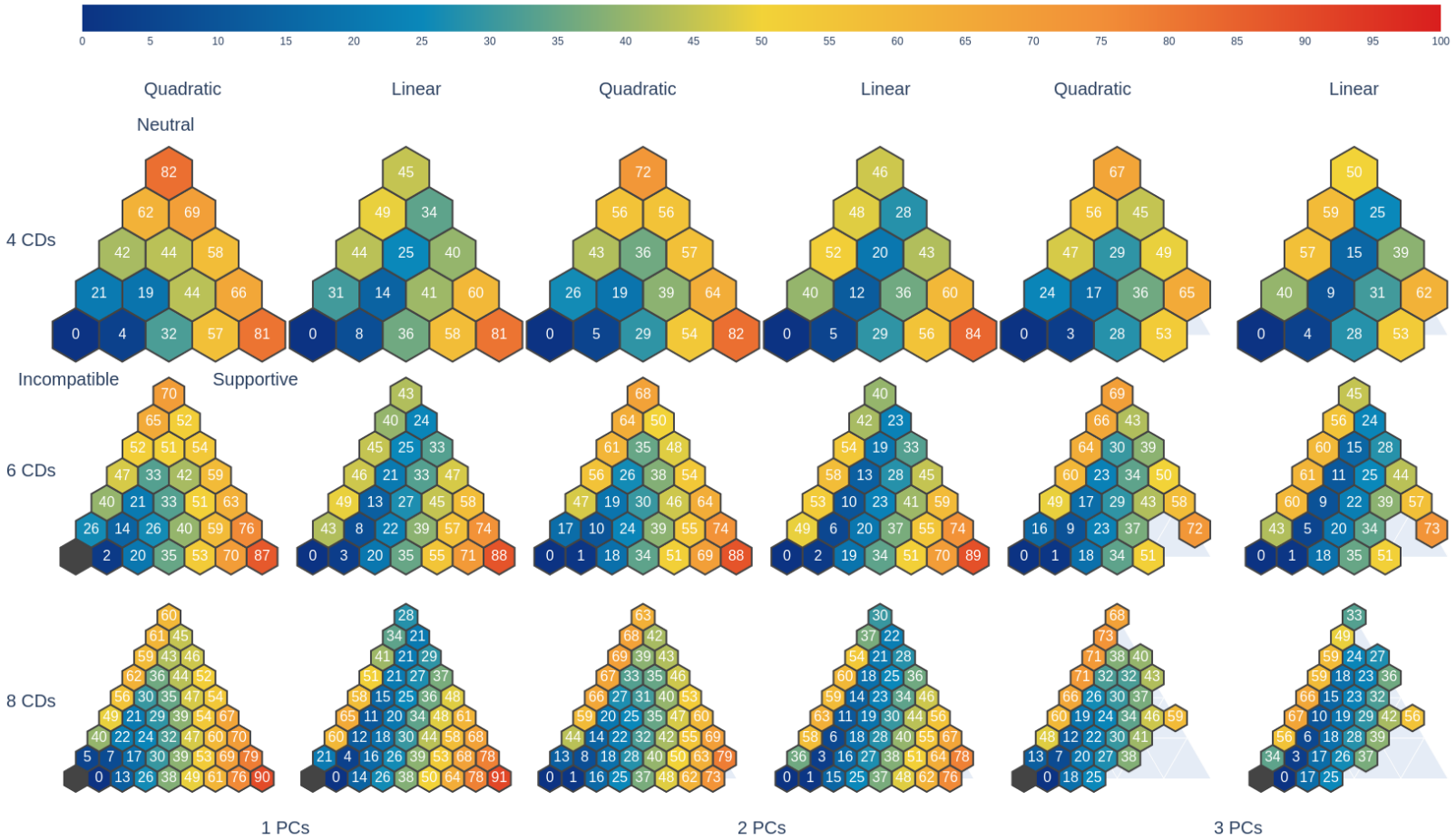

Figure 5.6: These ternary plots display the arithmetic averages of *option count* in the PC1-subsociety of the drawn tuples. The data is split up according to model variant, n_{CD} and n_{PC}.

SP1 Similar to the ternaries for the acceptance rate of PC1, the hexagons become hotter the closer one gets to the s-corner. That is, support connections to PC1 promote pluralism in the PC1-subsociety.

SP2 Additionally, however, there is also a hotspot on the 0s-isoline. (This hotspot is located somewhat differently for linear vs. quadratic models, see below.)

SP3 Another striking feature of all ternaries is that the 1s-isoline has a very low PC1-pluralism when compared to the other s-isolines. For example, have a look at the 0i3n1s hexagon ($n_{CD} = 4$) for any pluralism measure, model variant or n_{PC}. When compared to both the 1i3n0s and the 0i4n0s hexagon (i.e. exchange the support connection by either an incompatibility or a neutral connection, respectively), the 0i3n1s hexagon will have a *lower* PC1-pluralism. The same holds for all other 1s-hexagons. Thus, the 1s-isoline is an exeption to the above rule (SP1) that more s-connections to PC1 promote pluralism in the PC1-subsociety.

There are two main differences between the ternaries:

DP1 Even though the general heat pattern in the ternaries is remarkably similar for different pluralism measures, there is one significant difference: The hot area around the s-corner has a different 'size' depending on the pluralism measure. It is biggest for strength of the weak, second biggest for entropy and smallest for option count.

DP2 I already mentioned that the hotspot on the 0s-isoline is somewhat different for the quadratic and for the linear model variant. In the ternaries for the quadratic variant, the hottest hexagon is on the neutral corner and the 0s-hexagons become hotter with more n-connections. In the ternaries for the linear variant, however, the hottest 0s-hexagon is in between the i- and the n-corner. For example, have a look at the entropy ternary (linear variant, $n_{CD} = 8$, $n_{PC} = 1$). Here the hottest hexagon on the 0s-isoline is the 5i3n0s hexagon. This is just an example, similar results hold for the other measures and other values of n_{CD} and n_{PC}.

Potential Explanation

First of all, it might be surprising that the ternaries look so similar for different values of n_{CD}. After all, if there are more CDs available in the structure, we might expect there to be more pluralism in the fixpoints. And in some sense there probably is, but remember from the definition of the pluralism measures (sec. 3.4) that each measure was (in its own way) normalised with the maximum number of realisable options $\min(\{ n_{CD} + 1, n_{FP} \})$ where n_{FP} denotes the number of fixpoints in the relevant subsociety. Here, n_{FP} denotes the number of fixpoints accepting PC1. Thus, for sufficiently large PC1-subsocieties, the contribution of additional CDs in the structure is cancelled out by the normalisation of the measures. This explains why the ternaries look so similar for different values of n_{CD}.

Now, let's turn to the more substantial findings.

SP1 Support connections promote pluralism in the PC1-subsociety. This fact is explained separately for the pluralism measures:

Entropy and SoW: There are two factors that explain this finding for entropy and strength of the weak.

1. The *global* entropy (i.e. not in the PC1-subsociety, but among all fixpoints) is homogeneously high all over the ternaries, no matter the model variant, n_{CD} or n_{PC} (see the appendix F). This means that each CD can be expected to be accepted in roughly the same number of fixpoints per society.
2. Acceptance mechanism M1 is the main contributor to acceptance of PC1 for hexagons with one and more support connections. As a consequence, most FPs accepting PC1 (short: PC1-FPs) will also accept a supportive CD.

Now, suppose we are on the 2s-isoline. Most PC1-FPs will accept one of the two supportive CDs (due to the second point) and both of the two supportive CDs will be accepted a similar number of times (due to the first point). Of course, there might be a few PC1-FPs accepting a neutral CD (M2) or no CD (M3), but mostly they will be more or less evenly spread out over the two supportive CDs. Moving to the 3s-isoline, the same results holds, only

that the PC1-FPs will be mostly and evenly spread out over *three* supportive CDs instead of just two. Thus, with more support connections, the PC1-subsociety will be spread out evenly over more CDs. As a consequence, both entropy and strength of the weak in the PC1-subsociety increase. For entropy, this is plain, since entropy is a measure for how evenly spread out a distribution is over *all* CD-options (entropy is maximal iff the distribution is homogeneous). Strength of the weak, too, is somewhat sensitive to how evenly spread out a distribution is. If the percentage of FPs accepting the strongest option (a particular supportive CD) is roughly the same as the percentage of FPs accepting the other supportive CDs, then this percentage will decrease with more support connections, i.e. strength of the weak will increase.

OptCount: Option Count, on the other hand, is not at all sensitive to the distribution. All that counts is the total number of options realised at least once. Now, suppose M1 was the only mechanism leading to acceptance of PC1. Then with each additional supportive CD one more CD-Option would be realised at least once. It would be clear why option count increases with the number of support connections. But M1 isn't the only mechanism. What about M2 and M3? Why doesn't their contribution mess up this explanation? Remember from fig. 5.3 that the probability for M2 or M3 is relatively low, especially when moving away from the neutral corner. Since the societies in my study are rather small ($n_{FP} = 30$), perhaps the probability for M2 and M3 to contribute even a single non-supportive CD-option to the PC1-subsociety is comparatively low. As a consequence, option count increases with support connections. In bigger societies, we'd expect an entirely different picture, though. This is in line with the results in section 5.1.4 where we return to this issue.

SP2 There is a hotspot on the 0s-isoline. See explanation of DP2.

SP3 Let's postpone this discussion until after discussing DP2.

DP1 The size of the hotspot around the s-corner varies between the plural-

ism measures. I have no real explanation for this yet. Perhaps this is to be expected simply because the different measures work differently, including their normalisation which might play an important part in explaining the differences. Also, keep in mind that the ternaries for option count will look very different for bigger societies (see sec. 5.1.4).

DP2 The hotspot on the 0s-isoline is located differently for the quadratic and the linear model. In the quadratic model, it is located on the n-corner. In the linear model, it is located roughly in the middle of the isoline.

Quadratic: On the 0s-isoline, M1 plays no role, but M2 and M3 do. However, it's mostly M2 that contributes to acceptance of PC1. Again, we can assume that the fixpoints are distributed more or less evenly over the CDs. This means that on the n-corner most of the PC1-subsociety accepts neutral CDs (save for the few M3 fixpoints) and it is more or less evenly distributed over these neutral CDs. When moving towards the i-corner on the 0s-isoline, the PC1-fixpoints are distributed over less neutral CDs, resulting in a decrease of entropy and strength of the weak (this mirrors the explanation for SP1). Thus, we expect a hotspot of PC1-pluralism in the neutral corner and this is precisely what happens. For option count, the story is even more straightforward. With each neutral connection added to the structure, there is one more CD-option that can be realised with M2. Thus, option count increases towards the neutral corner. M3 plays a minor role for 0 and 1 incompatibility connection (the rest being neutral, of course). If at all this only contributes to the hotspot on the neutral corner.

Linear: For the linear model, it is *prima facie* not clear why there would be a hotspot at all on the 0s-isoline. After all, in the linear model M3 is the main mechanism leading to acceptance of PC1 on the 0s-isoline. (There might be a few occurrences of M2 around the neutral corner.) As a consequence, most fixpoints realise the CD-option of accepting no CD. Thus, we would expect little pluralism, no matter the measure. Nonetheless, the 0s-isolines have some pretty hot hexagons with an entropy of up to 70 in the 5i3n0s hexagon ($n_{PC} = 1$). What's happening here? I suspect that this has

something to do with the normalisation of the pluralism measures. On the 0s-isolines of the linear model we have very low acceptance rates, much lower than for the quadratic model. In the middle of the isoline (where the hotspot is typically located) we have an average acceptance rate of about 5-10%, so we can expect about two PC1-fixpoints per society (since there are 30 agents per society). Thus, the pluralism measures are normalised with $max = 2$ instead of $max = n_{CD} + 1$. This means that we will get a maximum pluralism score of 100 iff each of both fixpoints realises a different CD-option and 0 iff they realise the same option. (This holds for any pluralism measure due to their normalisation.) If there is only one fixpoint, no pluralism score is calculated and it does not affect the average pluralism scores displayed in the ternaries. For some reason, these tiny PC1-subsocieties realise the maximal pluralism score often enough to result in significant average values for pluralism. This is, of course, speculative. But if it's true then it seems that the present results (for the 0s-isoline, linear model) are *skewed* due to the small number of 30 agents per society resulting in tiny PC1-subsocieties on the 0s-isoline. For this reason, I have simulated the ternaries for $n_{CD} = 4$, $n_{PC} = 1$ once more with 300 instead of 30 agents per society (the rest of the study design being equal). I discuss these findings in the next section 5.1.4. As it stands, we should expect little pluralism on the 0s-isoline of the linear model. If at all, there should be a hotspot on the neutral corner, just like in the quadratic model, since that's where some M2-acceptance is possible in addition to M3-acceptance.

SP3 Now let's return to SP3: The 1s-isoline is significantly cooler compared to the others. I have explained above (SP1) why the 1s-isoline is cooler when compared to s-isolines with more than one support connection. But why is it cooler when compared to the 0s-isoline? I will here focus only on the *quadratic model* since we have seen that the high pluralism values of the linear model on the 0s-isoline are dubious. I discuss this point separately for the different pluralism measures.

Entropy and SoW: As we have seen, the hexagons close to the n-corner on the 0s-isoline have high pluralism (in the quadratic model), because the PC1-fixpoints are distributed more or less evenly over the neutral CDs. Now, if we move to the 1s-isoline, there is a significant jump in the acceptance rates, because the additional supportive CD contributes a lot of PC1-fixpoints via M1. However, these additional fixpoints all realise the same CD-option, namely the supportive CD. Thus, even though some of the fixpoints are evenly spread out over the neutral CDs, there is still a heavy emphasis on the supportive CD. In terms of entropy, the distribution is much less homogeneous, leading to a decrease. In terms of strength of the weak, there is now a strongest option (the supportive CD) with significantly more fixpoints than the others, leading to a decrease as well.

OptCount: For option count, however, this should not hold. Again, the distribution does not matter, only the number of options that are realised at least once. And this number should not decrease, at least not when comparing hexagons on the same n-isoline (i.e. when exchanging an incompatibility for a support connection, meaning that M2 can contribute as much as before). Nonetheless, the 1s-isoline is much cooler than the 0s-isoline. Again, I suspect that this has something to do with the normalisation of, in this case, option count. The acceptance rates are rather low on the 0s-isoline. They are higher for the quadratic model than for the linear model, but still typically less than 20%, i.e. less than 6 fixpoints. For $n_{CD} > 4$ this means that option count can be expected to be normalised with the number of fixpoints and not the number of CD-options. On the 1s-isoline, however, the acceptance rates are high enough such that option count is normalised with the number of CD-options (which is a bigger number). As a consequence, there is a higher probability that a bigger number appears in option count's denominator: The 1s-isoline is cooler than the 0s-isoline. But again, for bigger societies the PC1-pluralism ternaries for option count should look completely different.

Before wrapping up, let's have a look at larger societies.

5.1.4 Large Societies

In the discussion of the results for PC1-pluralism we have seen that there was quite some trouble due to the fact that the societies in my study have a size of only 30 agents. In particular, this has skewed the results for the 0s-isoline in the linear model and for option count in general. For this reason, I have run another setup of the study which is exactly the same, but with 300 instead of 30 initial commitments per structure. Of course, this requires much more computational power. For this reason, I only simulated societies with $n_{CD} = 4$, $n_{PC} = 1$. The results for the acceptance rate, split up into M1–3, are shown in fig. 5.7. The results for PC1-pluralism are shown in fig. 5.8. In what follows, I discuss both the 0s-isoline in the linear model as well as option count, because for both the results were skewed in small societies.

Let's start with the *0s-isoline in the linear model*, before turning to option count. In the last section, I argued that we would not expect any pluralism on this isoline. This is because M3 is the main contributor to PC1-acceptance and M3 does not by itself foster pluralism (since it's only realised with a particular CD-option) in contrast to M1 and M2. I hypothesized that it was the small acceptance rates, together with the normalisation of the pluralism measures, that lead to the hotspots on the 0s-isoline. It seems that this hypothesis is at least partially falsified by the present ternaries. Here, even the smallest average overall acceptance rate of 3% in the incompatibility corner (just add up the contributions of M1–3) corresponds to a PC1 subsociety of 9 agents. Thus, normalisation should not be an issue here, since $9 > 4 + 1$. But still, there is significant pluralism on this isoline for all pluralism measures. The reason for this appears to be that M2 contributes significantly even in the linear model (contrary to what figure 5.3 suggested): There is up to 6% M2-acceptance in the neutral corner. Since there are so many agents, we can expect these fixpoints to be more or less evenly spread out over the neutral CDs. Thus, even though there is 11% M3-acceptance (with only one option), this leads to significant pluralism values for entropy and strength of the weak (we'll discuss option count in a moment). However, remember from the last section that we can expect M2 to be weaker with increasing n_{PC}.

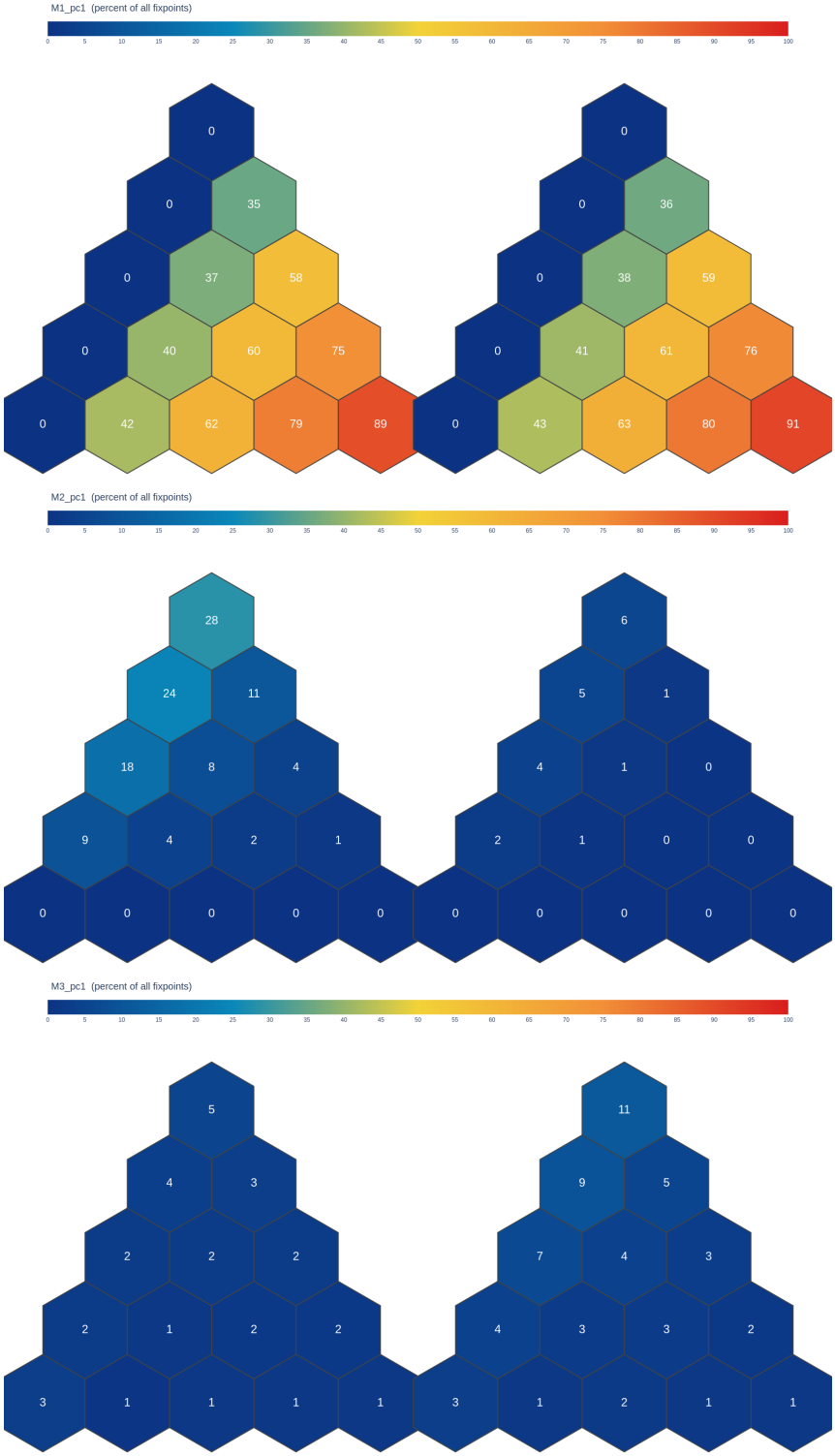

Figure 5.7: These three pairs of ternaries show the arithmetic averages for *M1-, M2- and M3-acceptance* (from top to bottom), in the *large* societies (300 ICs per structure) with $n_{CD} = 4$, $n_{PC} = 1$. The left of each pair shows the averages for the quadratic model, the right shows them for the linear model. Note that these numbers are just the acceptance rates split up. Thus, they don't sum to 100, but to the average overall acceptance rate in the respective hexagon.

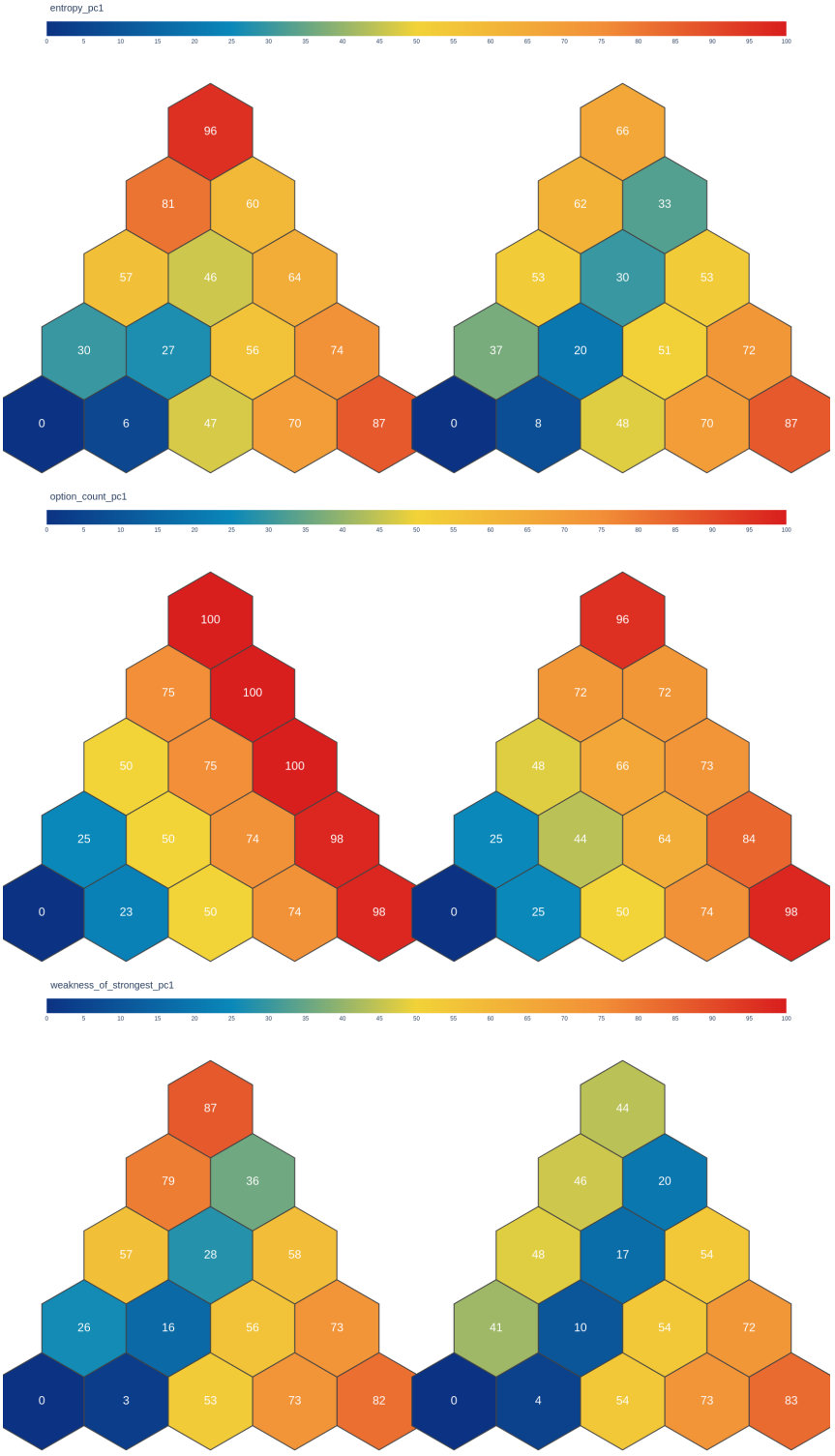

Figure 5.8: These three pairs of ternaries show the arithmetic averages for *entropy, option count and strength of the weak* (from top to bottom), in the *large* societies (300 ICs per structure) with $n_{CD} = 4$, $n_{PC} = 1$. The left of each pair shows the averages for the quadratic model, the right shows them for the linear model.

Since only M2 fosters pluralism, we can expect less pluralism (on the 0s-isoline) with more alternative political conceptions. Nonetheless, I conclude that the unexpected result from the last section, i.e. that in the linear model the hotspot on the 0s-isoline is *not* on the neutral corner even though that's where we'd expect one if at all, does not hold in the larger societies. Instead, the hottest hexagon is on the neutral corner. This is as expected, because M2 is strongest on the neutral corner. Thus, this implausible result was only due to the small society size of 30 agents.

Now, let's turn to *option count*. As I speculated in the last section, we see entirely different ternaries now. First, the 1s-isoline is not a cool spot anymore. Second, the number of support connections is not a decisive factor anymore. Instead, the main predictor for high PC1-pluralism is the absence of incompatibility connections to PC1. There is no big difference between support and neutral connections. The reason for this is simple: For option count it does not matter whether a CD-option is realised once or a hundred times. Even if there is only a comparatively slim chance for any particular process to realise M2, it will nonetheless happen at least once per society, because there are so many agents. As a consequence, there is a high probability that every CD that is either neutral or supportive will be accepted at least once together with PC1. Thus, it makes no difference for option count how many neutral vs. supportive connections there are. However, the number of incompatibility connections does, of course, make a difference, because an incompatible CD cannot be accepted together with PC1. Thus, every additional incompatible CD takes away one CD-option for the PC1-subsociety, leading to the decrease of PC1-pluralism when moving towards the i-corner. So far, so good. It should be noted, however, that there is a caveat regarding the linear model. As you can see on the i-isolines, there is a noticeable, though not huge, decrease in option count *in between* the neutral and the supportive end of the isolines. The reason for this is that M2 is not as strong a mechanism in the linear model as it is in the quadratic model. With more than 1 support connections, M2-acceptance drops to zero. In particular, it is unclear how things look in the linear model for $n_{PC} > 1$, because we can expect M2-acceptance to be even lower in these societies (see DA1 in section 5.1.2). Nonetheless, as of now, the implausible result for option count (i.e. a lower pluralism on the 1s-isoline when compared to the

0s-isoline) does not hold in large societies.

Thus, both implausible results from the last section vanish once we consider larger societies. Future studies will have to take this into account by either changing the normalisation or by simulating larger societies or both.

This concludes my presentation and discussion of the simulation results. Let's recap before turning to their interpretation.

- In the present study, *consensus* on a PC is mainly facilitated by CDs that support that PC. Besides that, CDs that are neutral about a PC also contribute somewhat to consensus on that PC, though the mechanism and extent of this depends on the model variant and on n_{PC}.
- Regarding *pluralism in the PC-subsociety*, the results of the study differ depending on the pluralism measure.
  - According to the two distribution-sensitive measures, entropy and strength of the weak, PC-pluralism is facilitated by two features of the dialectical structures: First, if there are *many* CDs that support a PC, then there will be pluralism in the PC-subsociety. Second, if *no* CD supports PC, but many are neutral about it, then there will also be pluralism in the PC-subsociety. However, if *exactly one* CD supports PC, then there will be little PC-pluralism.
  - According to the distribution-insensitive measure, option count, the decisive factor for PC-pluralism is the number of CDs that are incompatible with PC. If there are many such CDs, then there will be little pluralism. The numbers of neutral vs. support connections, on the other hand, don't make a huge difference.

Finally, it should be noted that there is a caveat regarding neutral connections in the linear model. Their positive influence on PC-pluralism depends on the mechanism M2 which is not particularly strong in the linear model. As of now, it looks like it's strong enough, but it is possible that these results are not robust for $n_{PC} > 1$.

On the bottom line, I submit that the results are by and large plausible, though we should take the results about neutral connections in the linear model with a grain of salt. Of course, we do not know to what extent the overall results will turn out to be robust once the formal model and the study

design are varied. But it does not seem like the results are wildly implausible or go against all expectations, at least when considering large societies as above. This is good, because it give us confidence that the formal model and study design do what they are supposed to, and that the results are not just some odd artefacts of the modeling approach but will actually inform us about the research question and hypotheses.

5.2 Testing the hypotheses

Thus, let's turn to this issue. How should we interpret these results with respect to the research question and hypotheses?

5.2.1 Potential local overlapping consensus

Let's start with the research hypotheses regarding potential local overlapping consensuses. Here is a restatement from section 2.2.6:

Hypotheses L If it's not the case that most comprehensive doctrines in the dialectical structure support PC, then

1. it is improbable that there is a potential local overlapping consensus on PC in the weak sense.

2. it is improbable that there is a potential local overlapping consensus on PC in the strong sense.

3. it is improbable that there is a potential local overlapping consensus on PC of grade $r \geq 0.5$.

What does this mean in terms of hexagons? How can we check these hypotheses by looking at the ternary heatmaps? A natural idea is to have a look at each hexagon satisfying the if-clause of the hypotheses and find out whether the different kinds of potential local overlapping consensuses are in fact improbable in this hexagon, i.e. less than half of the societies in the hexagon exhibit them. This is indeed what I will be doing. This gives us the following predictions that the hypotheses L1–3 make about the present data:

Predictions L For every hexagon in the study, if the number of support connections to PC1 is less or equal to $n_{CD}/2$, then

1. less than half of the societies in that hexagon exhibit a potential local overlapping consensus on PC1 in the weak sense.
2. less than half of the societies in that hexagon exhibit a potential local overlapping consensus on PC1 in the strong sense.
3. less than half of the societies in that hexagon exhibit a potential local overlapping consensus on PC1 of grade $r \geq 0.5$.

If we find any hexagon for which the if-clause is satisfied but the then-clause is violated, then the respective hypothesis is falsified. Let's start with L1, the hypothesis about potential local overlapping consensuses in the weak sense.

Potential Local OC in the weak sense

In the last section we discussed the arithmetic mean of PC1-pluralism in the ternaries. This was, I think, very useful for understanding how the different connection types influence PC1-pluralism in the respective societies. For the purpose of falsifying the present hypotheses, however, a slightly different way of representing the data is more useful. In particular, I think the *median* values for PC1-pluralism will be more informative. Figures 5.9–5.10 show these median values.

As you can see, these ternaries look pretty much like the ternaries for the arithmetic mean. Thus, their general description and explanation would mirror the discussion of the arithmetic mean values from the last section (save for specific values which sometimes differ, of course). However, the specific values for the median are more informative for falsifying the hypotheses. Let me explain.

Suppose there is a relatively high median for PC1-pluralism on the *neutral corner*. For example, in the drawn tuples for $n_{CD} = 6$, $n_{PC} = 2$, quadratic model variant, the median entropy in the PC1-subsociety is 77. That means, by definition of median, that half of the tuples have an entropy of 77 or higher while the other half have an entropy of 77 or lower. Thus, we *would*

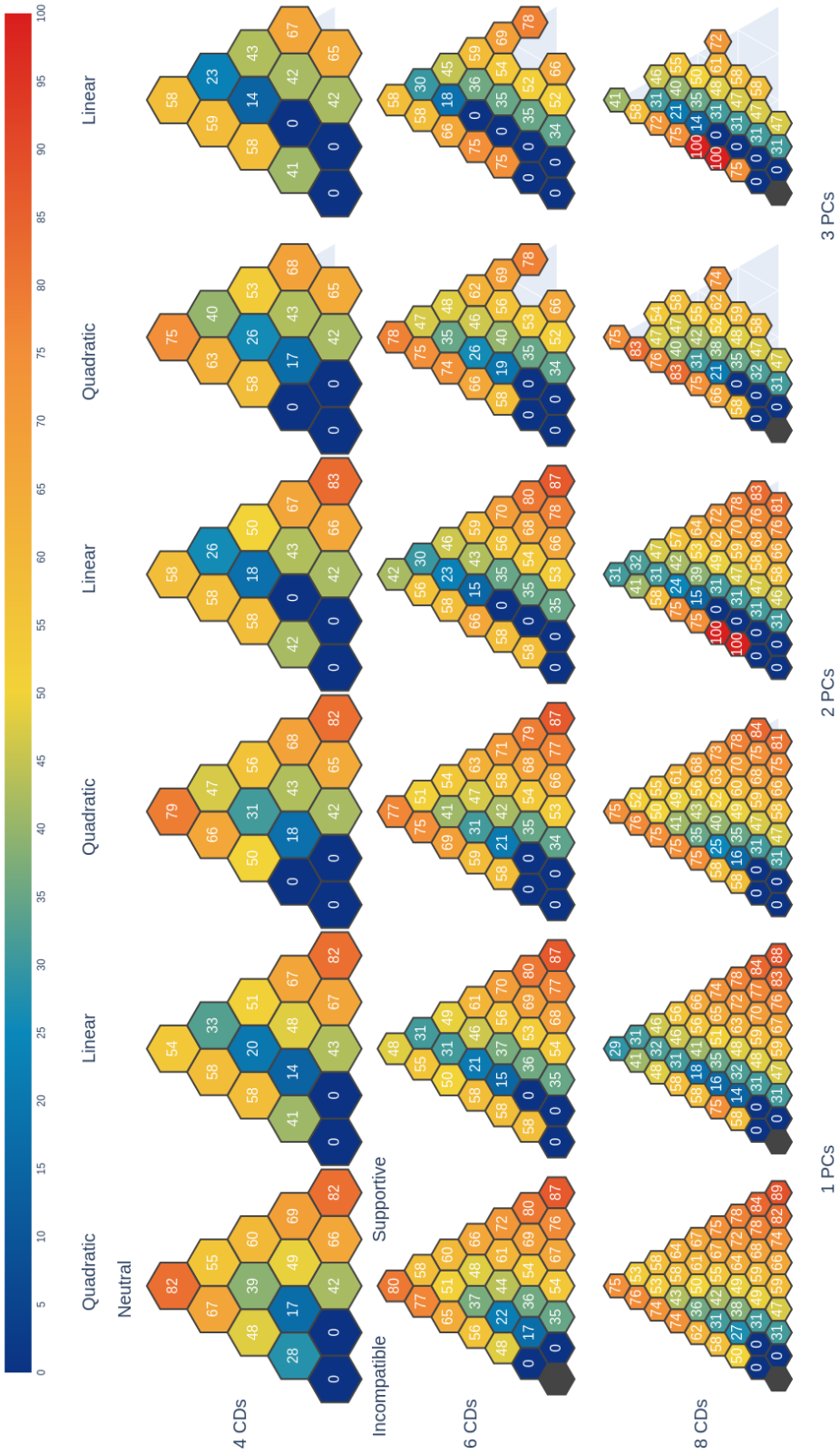

Figure 5.9: These ternary plots display the *median* of the *entropy* in the PC1-subsociety of the drawn tuples. The data is split up according to model variant, n_{CD} and n_{PC}.

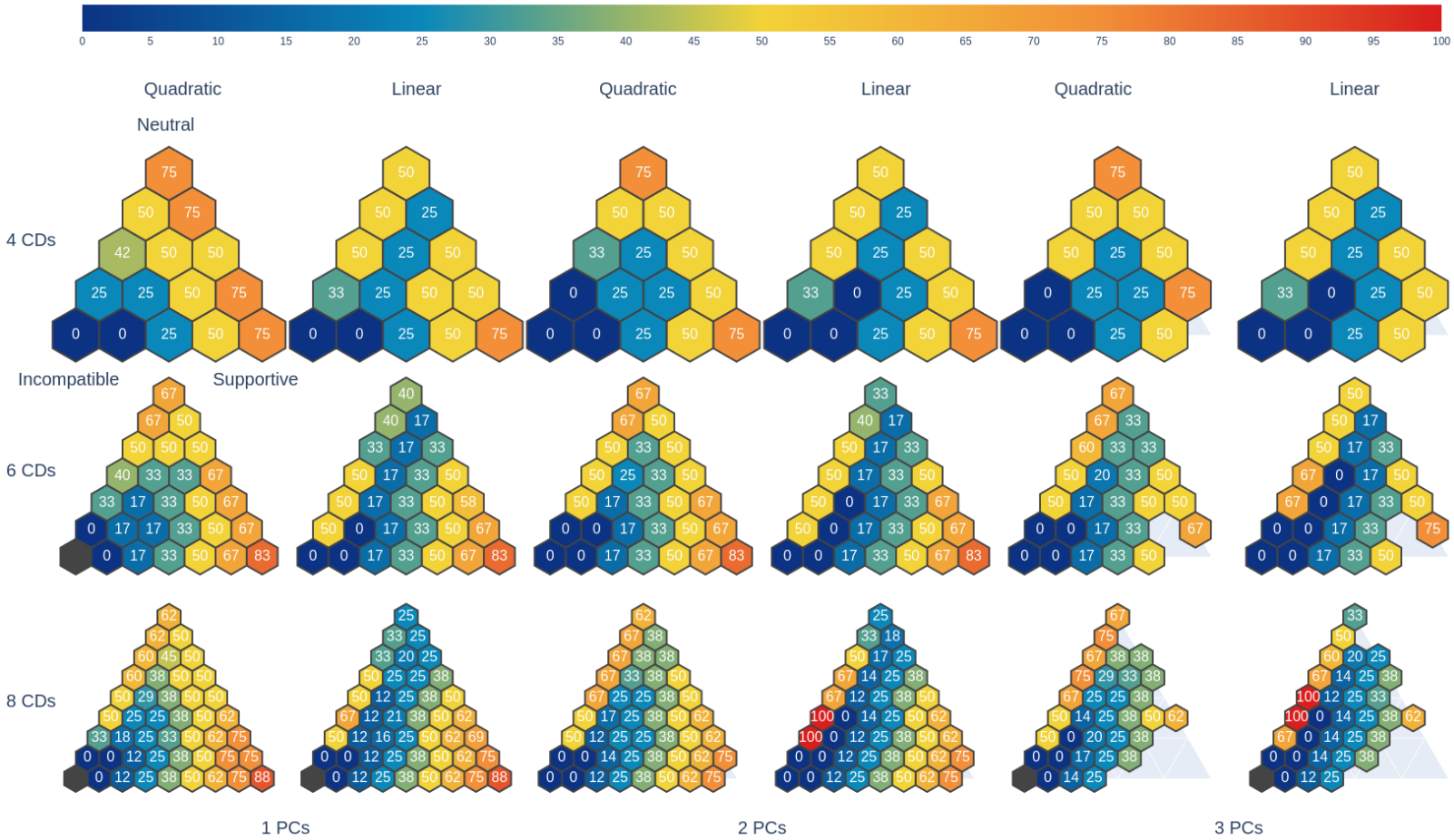

Figure 5.10: These ternary plots display the *median* of *option count* in the PC1-subsociety of the drawn tuples. The data is split up according to model variant, n_{CD} and n_{PC}.

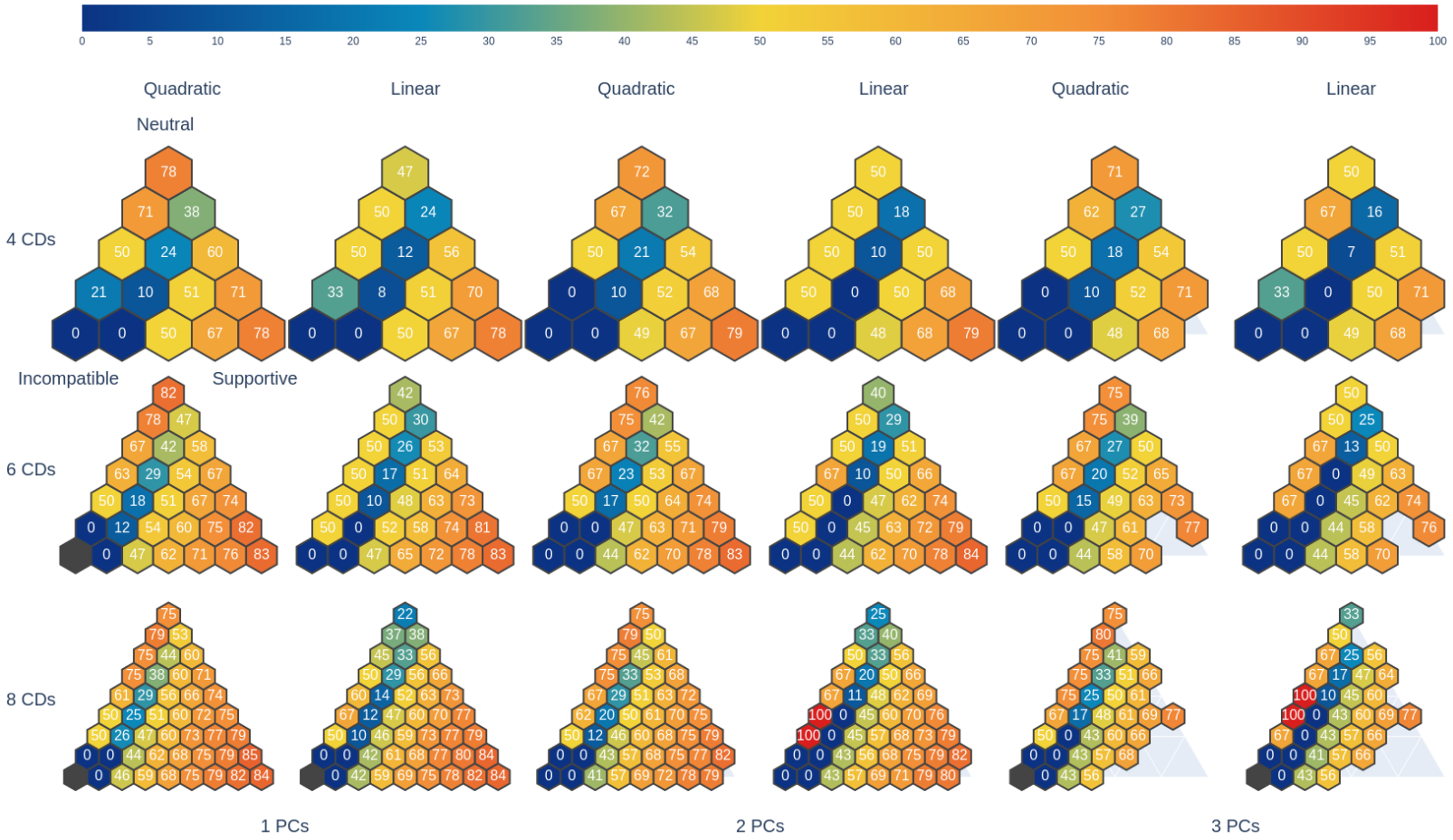

Figure 5.11: These ternary plots display the *median* of *strength of the weak* in the PC1-subsociety of the drawn tuples. The data is split up according to model variant, n_{CD} and n_{PC}.

have to set the pluralism threshold for entropy to 77 such that at least 50% of the drawn tuples count as pluralist. As a consequence, if it is plausible to say that the 'real' threshold is ≤ 77, then we can infer that at least 50% of the drawn tuples are pluralist. Since we draw exactly one tuple for each society, it follows that in at least 50% of the societies in this hexagon there is at least one tuple with pluralism in the PC1-subsociety, namely the one that was drawn. Thus, at least 50% of the societies in this hexagon exhibit a local overlapping consensus in the weak sense. Or, put in terms of probabilities: If we know nothing about a society besides that it falls into the 0i6n0s hexagon, then there is a probability $\geq 50\%$ that there is a potential local overlapping consensus in the weak sense in that society. As a consequence, the prediction of L1 is falsified by this hexagon. In essence, the strategy I suggest for falsifying L1 is to check the median values for PC1-pluralism in different hexagons. Whenever we find a hexagon where this median value is at least as high as the lowest plausible threshold for pluralism, we can infer that there is a probability of $\geq 50\%$ for a society in this hexagon to have at least one tuple with PC1-pluralism. If there is such a hexagon outside of the area around the support corner, i.e. a hexagon where it is not the case that most CDs support PC1, then hypothesis L1 is falsified.

Now, the question is, of course, whether there are such hexagons. For entropy, this seems to be the case, at least for the quadratic model. Here the median *entropy* in the PC1-subsociety of the drawn tuples in the neutral corner is between 75 and 82. This seems high enough to crack the lowest plausible threshold for pluralism (measured as entropy). In particular, it is not considerably lower than for the hexagons around the support corner. For *option count* in the quadratic model, the median value is in between 62 ($n_{CD} = 8,\ n_{PC} = 2$) and 75 ($n_{CD} = 4,\ n_{PC} = 3$). For the lower bound of 62, this means that in 50% of all drawn tuples, at least 5 out of 9 CD-otions are realised at least once in the PC1-subsociety. Again, this seems high enough to crack the lowest plausible threshold for pluralism (measured as option count). Regarding *strength of the weak* in the quadratic model, the median values are in between 71 ($n_{CD} = 4,\ n_{PC} = 3$) and 82 ($n_{CD} = 6,\ n_{PC} = 1$). For the lower bound, this means that in the median tuple the weak options (i.e. the ones not realising the strongest option) taken together have 71% of their theoretically possible strength of 80%, i.e. about 57% of fixpoints do not

realise the strongest CD-option. Again, this seems pluralist enough.

As a consequence, I submit that for all pluralism measures and all ternaries of the quadratic model variant, hypothesis L1 is falsified. That is, even if it's not the case that most CDs support PC1, there are hexagons such that a potential local overlapping consensus in the weak sense is probable. But what about the linear model variant? Here, too, we have significant median pluralism outside of the area around the support corner. In particular, there are such values in the 0s-isoline. However, in the last section we have seen that the small society size in my original study has skewed these results due to the normalisation of the pluralism measures. So what about large societies? Take a look at figure 5.12, where I have plotted the ternaries from section 5.1.4 with median values instead of the arithmetic mean.

It seems that in the linear model we have significant pluralism values on the 0s-isoline for all measures. With option count we get a score of 100 in the neutral corner. That's enough. With entropy we get 65 in the neutral corner. That's less than the lowest value in the quadratic model for the small societies, but still significant. And, I submit, it should be enough to count as pluralist (though this might be a matter of contention and require further argument). Finally, with strength of the weak we get a median score of 46 in the 1i3n0s- and 2i2n0s-hexagons. That is, the weak have 46% of their theoretically possible strength (80% for $n_{CD} = 4$), i.e. about 37% of fixpoints in the PC1-subsociety do not realise the strongest CD-option. This means that the strongest CD-option is realised in more that 50% of the cases. It is not entirely clear whether we should count such a tuple as pluralist. I am inclined to do so, but it might be a matter of contention. Let's say it does count as pluralist, though it is probably close to the lowest possible threshold for pluralism. If this characterisation is legitimate, then it seems that hypothesis L1 is also falsified for the linear model. However, there is still a caveat, because we have not seen the ternaries for large societies with $n_{PC} > 1$, just like in the last section.

All things considered, I submit that the present data falsifies hypothesis L1. This is particularly clear for the quadratic model. For the linear model this holds at least for option count, i.e. the most minimalist pluralism measure. Arguing that it also holds for entropy and strength of the weak (in the linear model) might require further careful argument concerning the

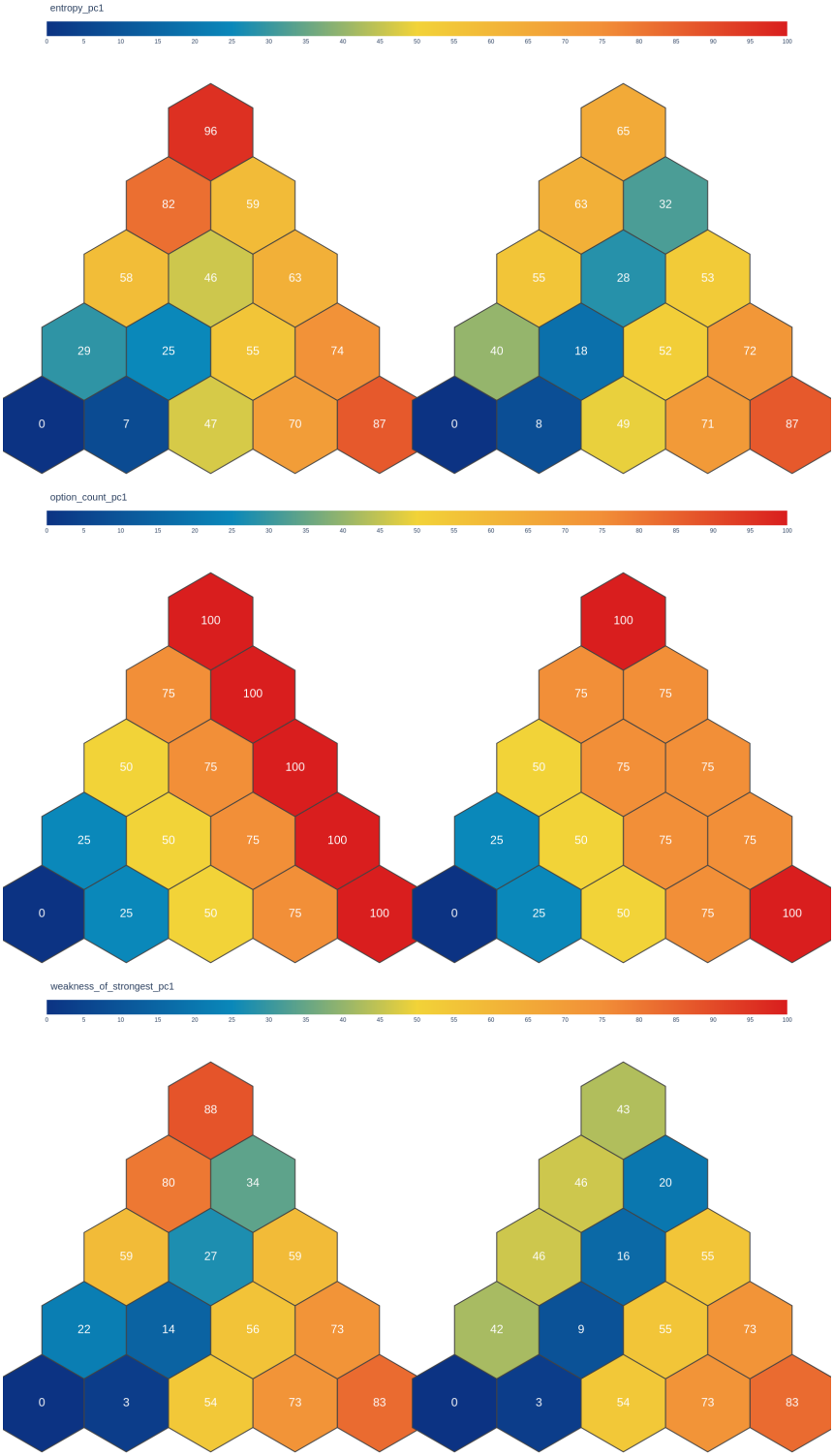

Figure 5.12: These three pairs of ternaries show the *median* values for *entropy, option count and strength of the weak* (from top to bottom), in the *large* societies (300 ICs per structure) with $n_{CD} = 4$, $n_{PC} = 1$. The left of each pair shows the averages for the quadratic model, the right shows them for the linear model.

pluralism thresholds for these measures.

Potential Local OC in the strong sense

Let's turn to hypothesis L2 stating that if it's not the case that most CDs support PC, then a potential local overlapping consensus in the strong sense is improbable. Here, the approach we used for falsifying L1 does not work, because it does not suffice that for most societies there is at least one tuple with pluralism in the PC1-subsociety. We need *all* tuples to exhibit such pluralism. Plainly, whether and how often this occurs is not an information we have access to until we simulate societies with branching, i.e. for each agent we generate all outcomes of LocalMRE, not just one, so that we can calculate PC1-pluralism in the whole space of justified belief systems for a given society (and not just one point in this space as we did in the present study). Even if the median pluralism value in some hexagon is really high, it is still possible that in most societies there is at least one tuple without pluralism. Thus, as of now, we cannot falsify L2.

Potential local OC of high grade

Let's turn to hypothesis L3 stating that if it's not the case that most comprehensive doctrines in the structure support PC, then a potential local overlapping consensus of grade $r \geq 0.5$ is improbable.

Since for each society in my study we only draw one random tuple from that society's space of justified belief systems, we cannot say much about the grade of a given society's potential local overlapping consensus. Nonetheless, we might be able to make a statistical inference about the arithmetic mean grade in a given hexagon. In particular, I will outline an argument concluding that the arithmetic mean grade of the potential local overlapping consensuses of societies in the neutral corners is $\bar{r} \geq 0.5$. If sound, this argument would seem to falsify L3's prediction that it is improbable for such societies to have an potential local overlapping consensus of grade $r \geq 0.5$. However, the argument requires a non-trivial *indifference assumption* about the algorithm LocalMRE that is currently unverified. Thus, L3's falsification is something like a *conditional* result. For ease of exposition, I will first sketch the argument using what I call the *homogeneity assumption* which is an overly

strong version of the required indifference assumption. Later I argue that we can relax it.

The basic idea goes like this. Suppose we know the 'real' entropy threshold for pluralism. Or, at least, we have a solid argument for a specific value to be the lowest plausible threshold. For each simulated society we could then mark the drawn tuple as either pluralist (1) or non-pluralist (0). For any hexagon, we can then calculate the *proportion* of tuples that are pluralist by calculating the arithmetic mean of the 0- and 1-markers. These proportions could be plotted in the familiar ternaries.

Now, the interesting point here is that this procedure (drawing a random tuple for each society in a hexagon and calculating the proportion of pluralist tuples in the hexagon) is a *random experiment*. And we can calculate an expected value for the outcome of the experiment, i.e. for the proportion of pluralist tuples in the hexagon. Interestingly, given the homogeneity assumption below, the expected proportion of pluralist tuples in the hexagon is identical to the arithmetic mean grade $\bar{r}$ of the societies in the hexagon. And, as always, if the sample size is large enough, we may assume that the measured value of the random experiment (the measured proportion of pluralist tuples in the hexagon) is close the expected value (the expected proportion of pluralist tuples in the hexagon). (In this case, the sample size is 50 times the number of heads in the given hexagon.) But since the expected value is *identical* to the arithmetic mean grade $\bar{r}$ of the societies in the hexagon, it follows likewise that the measured proportion of pluralist tuples in the hexagon is close to $\bar{r}$ in the hexagon. Thus, we can statistically infer $\bar{r}$ from the measured proportion of pluralist tuples in the hexagon.

I want to stress that this is nothing but standard procedure with an extra step attached to it. Suppose you want to figure out the bias of a coin. You might toss it 100 times and record the outcome, say, 70 times heads and 30 times tails. What does that tell you about the coin's bias? Well, you will assume that the measured value is close to the expected value. And the bias giving you an expected value of 70 times heads and 30 times tails is precisely 70:30. Thus, you infer that the coin's bias is 70:30 or close to that. I do basically the same but with an extra inference attached to it. Suppose that there are boundary conditions such that one can infer the coin's mass distribution from its bias. Then tossing the coin 100 times informs you not

only about its bias, but also about its mass distribution. In my case, drawing a random tuple per society and measuring the proportion of pluralist tuples in the hexagon informs me not only about the expected proportion, but also about $\bar{r}$.

The only thing left to do is to establish that, given the homogeneity assumption, the expected value for the proportion of pluralist tuples is in fact identical to $\bar{r}$ for any given hexagon. For each hexagon we have a number of heads $n_{heads} \geq 1$. For each of these heads we have 50 societies. For each of these 50 societies we draw one tuple from that society's space of justified belief systems. The homogeneity assumption states that the algorithm LocalMRE (for either quadratic or linear achievement function) samples the space of justified belief systems with a homogeneous probability distribution. That is, each tuple in the space of justified belief systems has an equal chance to be drawn. Let r_i be the grade of the potential local overlapping consensus in the ith society of the given hexagon (with $i = 1, \ldots, 50 \cdot n_{heads}$). This grade is defined as the proportion of tuples (in that society's space of justified belief systems) with a pluralism of CDs in the PC1-subsociety, see explication 4. Thus, if we draw a random tuple from the ith society's space of justified belief systems (with homogeneous probability distribution), then that tuple's *expected pluralism value* is $1 \cdot r_i + 0 \cdot (1 - r_i) = r_i$.

How can we use this to calculate the expected proportion of pluralist tuples in a given hexagon? For each society in a hexagon we draw a random tuple from that society's space of justified belief systems. As noted above, the proportion of pluralist tuples in the given hexagon is equal to the arithmetic mean of the pluralism values of the tuples in that hexagon. Now, since the expected value operator is linear, we can calculate the *expected proportion* of pluralist tuples in the hexagon by simply calculating the arithmetic mean of the *expected pluralism values* of the tuples in the hexagon. That is, the expected value for the proportion of pluralist tuples in the hexagon is $\frac{1}{50 \cdot n_{heads}} \sum_i r_i$. This is plainly identical to the mean grade $\bar{r}$ of the societies in the hexagon. Thus, the expected proportion of pluralist tuples in a hexagon is equal to $\bar{r}$. As a consequence, as promised above, we can measure the proportion of pluralist tuples in a hexagon and statistically infer the arithemic mean grade $\bar{r}$ to be close to that measured value. However, this reasoning relied on the strong homogeneity assumption that LocalMRE samples the space of justified belief

systems in a society with homogeneous probability distribution.

Suppose we have followed this procedure and, for example, the hexagon on the neutral corner can be inferred to have an average grade of $\bar{r} \geq 0.5$. What does that tell us about hypothesis L3? The hypothesis states that if it's not the case that most comprehensive doctrines support PC, then it is improbable that there is a potential local overlapping consensus of grade $r \geq 0.5$. It seems that in this case the hypothesis is falsified by the data.

Now, this procedure presupposes that we know the 'real' entropy threshold or at least have a solid argument for the lowest plausible threshold. But I don't. Still, we can rely on the familiar trick using the median ternaries. For each hexagon, the median value for PC1-pluralism gives us the threshold we *would* have to set such that exactly 50% of the drawn tuples are pluralist. Thus, it gives us the threshold we would have to set such that we can infer that $\bar{r} = 0.5$. As a consequence, if it is plausible to say that the median value of a given hexagon is higher than the lowest plausible pluralism threshold, then we can infer that even more of the drawn tuples are pluralist, i.e. we can infer that $\bar{r} \geq 0.5$. And this, too, falsifies hypothesis L3 if the hexagon is not from around the support corner, e.g. if the hexagon is on the neutral corner.

Thus, the falsification conditions for hypothesis L3 would be identical to the ones of L1, but only if the homogeneity assumption were to hold. Above, I have concluded that the present data falsifies L1 (without extra assumption), thus, it would also falsify L3 (given the extra assumption). However, the homogeneity assumption, i.e. that LocalMRE samples the space of justified belief systems with homogeneous probability distribution, is not generally true. As Freivogel and Cacean (2023) have shown, given the starting point of an individual agent (a set of initial commitments and a dialectical structure) it's not the case that all fixpoints that can result from an application of LocalMRE always have an equal chance to be drawn during such a process. For example, the following can be the case: Right in the beginning of the process for agent a_i, a random choice between two options is made. If the first option is chosen, then there is only one fixpoint that can result from the process. If the second option is chosen, then there are ten (other) fixpoints that can result from this process. In such a situation, the one fixpoint resulting from the first option will have a higher probability to be chosen

during equilibration than the other ten from the second option. (Due to this complication, Freivogel and Cacean distinguish the "process perspective", which is sensitive to these inhomogeneities, and the "result perspective", which is not.) Now, when simulating the whole society including agent a_i, tuples featuring the fixpoint from the first option (for agent a_i) are more likely to be drawn than tuples featuring one of the other ten fixpoints (for agent a_i). Thus, it is not generally the case that LocalMRE samples the space of justified belief systems with homogeneous probability distribution. Given a particular society, it might be that tuples with a high PC1-pluralism have a higher (or lower) chance to be drawn, because these tuples happen to contain fixpoints that are more likely (or less likely) to be drawn.

However, this is not necessarily a problem for the present analysis. If in some societies tuples with PC1-pluralism have a higher chance to be drawn, but in some other societies such tuples have a lower chance to be drawn, then these differences may average out. The crucial question is whether the algorithm *systematically* favours tuples with high PC1-pluralism. If this is the case, then we may not draw the inferences detailed above. We then have to assume that even if the median PC1-pluralism in some hexagon is pretty high, this might just be due to the algorithm favouring tuples with high PC1-pluralism. The arithmetic mean grade $\bar{r}$ may be much lower. In order to learn something about $\bar{r}$ in specific hexagons, we may then not get around simulating societies *with branching*, i.e. for each agent every possible fixpoint is calculated. This is extremely costly in terms of computational power and would hinder progress on the present questions.

Suppose, on the other hand, that LocalMRE does *not* systematically favour tuples with high PC1-pluralism, because it is on average indifferent about PC1-pluralism, i.e. it does not favour tuples with high or low PC1-pluralism. Call this the *indifference assumption*. In this case, things look better. Even though it is not generally the case that in the ith society of a given hexagon, the expected pluralism value of a drawn tuple is $1 \cdot r_i + 0 \cdot (1 - r_i) = r_i$ (since we relaxed the overly strong homogeneity assumption), the deviations from this value average out when calculating the expected arithmetic mean of these pluralism values for the whole hexagon, i.e. when calculating the expected proportion of pluralist tuples in the hexagon. Thus, in this case we may still assume that the expected value for the proportion of pluralist

tuples in the hexagon is close to $\frac{1}{50 \cdot n_{heads}} \sum_i r_i$ which is identical to $\bar{r}$.

If this is indeed the case, then we can draw interesting conclusions about potential local overlapping consensus of high grade *without* simulating all branches for every agent in every society. In particular, the reasoning given above is legitimate and hypothesis L3 is falsified. Testing the indifference assumption is crucial, not only for analysing the present data, but also for setting up subsequent studies in a computationally frugal way. If the indifference assumption is true, then we may learn much about the grades of potential overlapping consensus without simulating all branches for all agents. There are many ways to study whether LocalMRE systematically favours pluralist tuples, but one very straightforward way would be to simulate some hexagons *with* branching, then directly *calculating* the arithmetic mean grade $\bar{r}$ (instead of inferring it) and comparing the results with the proportion of pluralist tuples one would get without branching. If the results are similar, it gives us confidence that the indifference assumption holds and we may analyse the data accordingly for other hexagons as well. Perhaps there are better ways to study the indifference assumption, but this will not be my concern here.

A last remark: You may have noticed that I left out a possible outcome of studying the indifference assumption. What if LocalMRE on average does *neither* favour tuples with high PC1-pluralism *nor* is it indifferent to PC-pluralism, but instead it favours tuples with *low* PC1-pluralism? What follows from this will in general depend on the study and the research hypotheses. But, interestingly, in the present context it would follow that L3 is still falsified. If LocalMRE favours tuples with low pluralism, we may assume that the average grade $\bar{r}$ of some given hexagon is even higher than the measured proportion of pluralist tuples. Thus, in the present context we only require a weaker version of the indifference assumption, i.e. we only require that $\bar{r}$ is greater than or equal to the expected proportion of pluralist tuples.

Let's wrap up the considerations about hypothesis L3. There are two cases:

1. It is *not* true that LocalMRE on average favours tuples with high PC1-pluralism: If a hexagon has a median pluralism value that is higher than the lowest plausible threshold for pluralism, then we can infer that

the average grade $\bar{r}$ of the societies in that hexagon is ≥ 0.5. If there is such a hexagon outside of the area around the support corner, e.g. on the neutral corner, then hypothesis L3 is falsified. This mirrors the falsification conditions I suggested for L1. Since L1 was by and large falsified by the data, I submit that L3 is falsified as well. (Analogously, there is a remaining caveat for the linear model with $n_{PC} > 1$.)

2. It *is* true that LocalMRE on average favours tuples with high PC1-pluralism: We cannot draw the inference detailed above. In order to learn something about the grade of potential local overlapping consensus in different hexagons, we must simulate these societies with branching, i.e. for every agent all fixpoints are calculated.

We should wait for some studies on how LocalMRE draws tuples. This information will get us a long way to interpreting the present data.

5.2.2 Potential global overlapping consensus

Let's turn to the hypotheses about potential global overlapping consensuses:

Hypotheses G If it's not the case that most comprehensive doctrines in the dialectical structure support PC, then

1. it is improbable that there is a potential global overlapping consensus on PC in the weak sense.

2. it is improbable that there is a potential global overlapping consensus on PC in the strong sense.

3. it is improbable that there is a potential global overlapping consensus on PC of grade $r \geq 0.5$.

In analogy to the considerations above about the local variants of the research hypotheses, these global variants yield the following predictions:

Predictions G For every hexagon in the study, if the number of support connections to PC1 is less or equal to $n_{CD}/2$, then

1. less than half of the societies in that hexagon exhibit a potential global overlapping consensus on PC1 in the weak sense.

2. less than half of the societies in that hexagon exhibit a potential global overlapping consensus on PC1 in the strong sense.

3. less than half of the societies in that hexagon exhibit a potential global overlapping consensus on PC1 of grade $r \geq 0.5$.

In the explications 4 of the local kinds of overlapping consensuses, the tuples need to exhibit only PC-pluralism. In the explications 3 of the global kinds of overlapping consensuses, on the other hand, the tuples need to exhibit not only PC-pluralism, but also consensus on PC. In the acceptance rate ternaries of the last section (displaying arithmetic means), we have seen that significant consensus occurs only in the area around the support corner. There is also some consensus on the neutral corner, but the acceptance rates are here always well below 50%. Thus, the limiting factor for the different kinds of a potential global overlapping consensus seems to be consensus and not pluralism. The only area in the ternaries with both high average consensus and high average pluralism is the area around the support corner.

As a consequence, we cannot try to falsify hypotheses G using the same strategy that I used when trying to falsify hypotheses L. It seems that outside of the support area there just aren't many tuples with consensus, perhaps there are none at all. As of now, I cannot think of an alternative strategy for falsifying hypotheses G. Thus, I assume that the present data does not warrant falsifying them.

But perhaps we can say more than that. Perhaps we cannot only *not* show that the consequents of the hypotheses are false, but we *can* show that the consequents are true. Of course, this wouldn't verify the hypotheses in a strong sense (as Popper warns us), but it would show that the hypotheses make correct predictions for the present data. And this, in turn, would contribute to answering the general research question.

For this task it will again be useful to have a look at the median values instead of the arithmetic mean. In figure 5.13 you see the median acceptance rate ternaries. Again, the ternaries look very similar to the ones for the arithmetic means, thus, their description and explanation mirrors the one in section 5.1.2 and I will not repeat these findings.

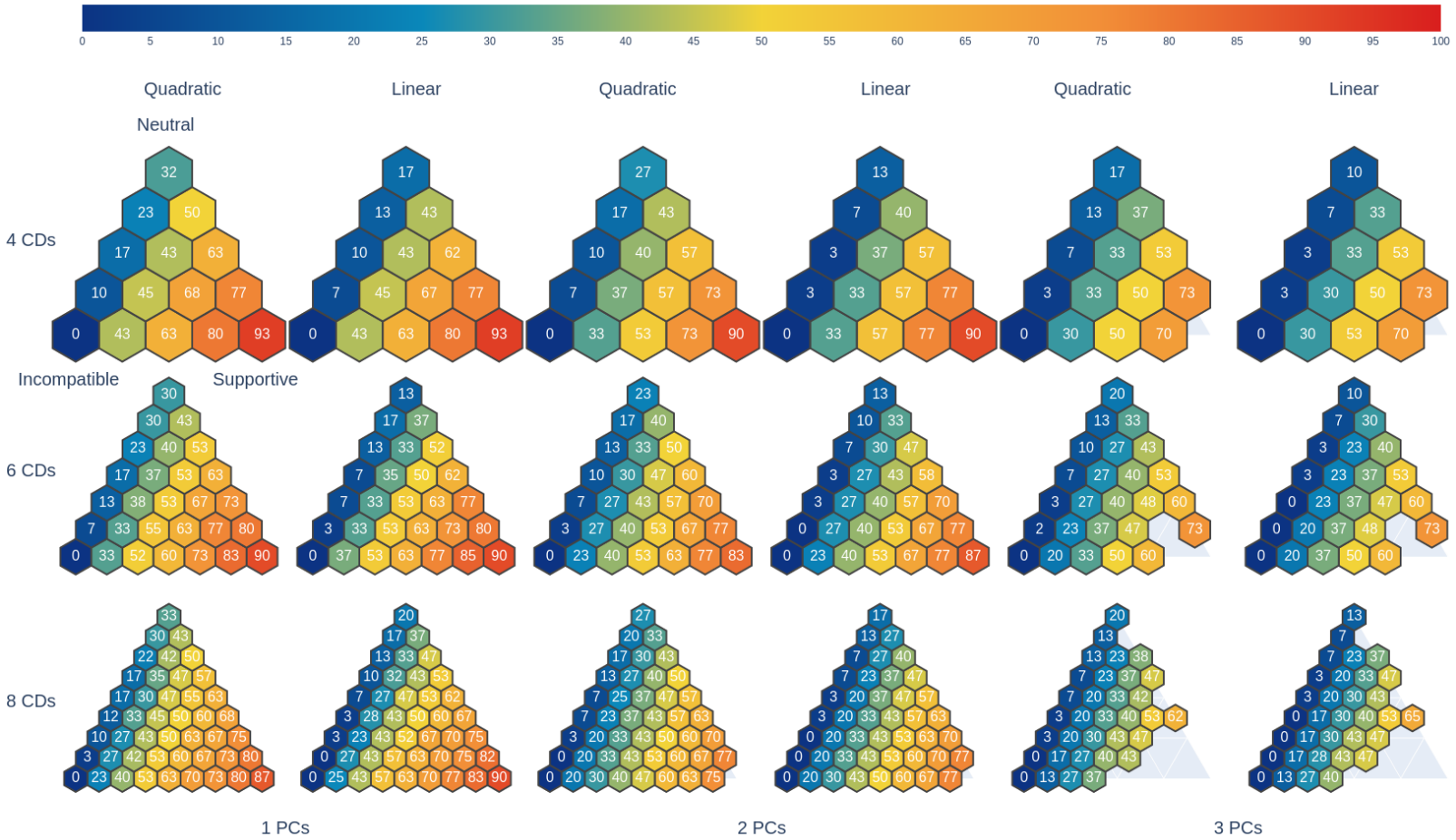

Figure 5.13: These ternary plots display the *median* values of the *acceptance rates*. The data is split up according to model variant, n_{CD} and n_{PC}.

Potential global OC in the weak sense

In order to verify the consequent of hypothesis G1, we need to show that outside of the support area it is improbable that there is even one tuple with consensus and pluralism. Given the limited data we have on the tuples (only one tuple drawn per society), this seems impossible to show. In a sense, this mirrors the difficulty of falsifying hypothesis L2 about the potential local OC in the strong sense, i.e. showing that it is probable that all tuples exhibit PC1-pluralism: Even if the median pluralism is very high on the neutral corner, it is still possible that there is a non-pluralist tuple in all (or most) societies. Here, the situation is similar. Even if the median consensus is very low outside the support area, it is still possible that there is a hexagon with at least one tuple with consensus and pluralism in all (or most) societies. Thus, the present data does not allow us to say that the consequent of G1 is true in hexagons outside the support area.

Potential global OC in the strong sense

In order to verify the consequent of hypothesis G2, we need to show that in each hexagon outside of the support area, it is improbable that *all* tuples of a society exhibit consensus and pluralism. And it seems we can do so by using the familiar strategy of analysing the median ternaries, this time, we'll focus on median acceptance rates.

As fig. 5.13 shows, if half (or less) of the comprehensive doctrines support PC1, then the median acceptance rates are invariably lower than 70%. In fact, for $n_{PC} > 1$ they are always lower than 60%. This is arguably not enough for consensus. Again, the median gives us the threshold that we would have to set such that exactly 50% of the drawn tuples crack it. Thus, if the median values are not enough for consensus, we can say that outside of the support area, *at least* 50% of the drawn tuples do not exhibit consensus and, *a fortiori*, do not exhibit consensus and pluralism. As a consequence, at least 50% of the societies in a given hexagon have at least one tuple without consensus and pluralism (namely the one that was drawn). Therefore, in at least 50% of the societies in a given hexagon, there is no potential global OC in the strong sense. The prediction of hypothesis L2 is true in the present data: Outside the support area, a potential global OC in the strong sense is improbable.

Potential global OC of high grade

Lastly, let's turn to hypothesis L3. Again, we cannot falsify it, but we might be able to say that its prediction holds in the present data. The strategy is similar to the one I used when trying to figure out the average grade $\bar{r}$ of the potential local OCs in the societies outside the support area.

Again, we need a form of the indifference assumption. This time, the indifference assumption is not that LocalMRE, when sampling spaces of justified belief systems, is on average indifferent to whether tuples exhibit PC1-pluralism. Instead, the indifference assumption is that LocalMRE, when sampling spaces of justified belief systems, is on average indifferent to whether tuples exhibit PC1-pluralism *and* consensus on PC1. If this holds, then in a given hexagon the expected proportion of tuples with pluralism *and* consensus is close to the arithmetic mean grade $\bar{r}$ of the potential *global* overlapping consensus in the societies of that hexagon. (The reasoning is completely analogous to the one given in the last section.)

Thus, we need to distinguish two cases again:

First, it *is* the case that LocalMRE is on average indifferent to whether tuples exhibit pluralism and consensus. In the discussion about the potential global OC in the strong sense we have already seen that in any hexagon outside the support area, at least 50% of tuples do not exhibit consensus and pluralism. Given the indifference assumption, the expected value of the proportion of tuples with consensus and pluralism is identical to the mean grade $\bar{r}$ of the potential global OC in the societies. Thus, we can infer that $\bar{r} \leq 0.5$ in hexagons outside the support corner. This verifies G3's prediction that outside the support area a potential local overlapping consensus of grade $r \geq 0.5$ is improbable. Likewise, if LocalMRE is on average not indifferent, but favours tuples with pluralism and consensus, then we can assume that $\bar{r}$ is even lower, i.e. G3's prediction is verified as well. (This latter point is again symmetrical to the considerations about falsifying L3.)

Second, it is *not* the case that LocalMRE is on average indifferent to whether tuples exhibit pluralism and consensus, because it favours tuples that do not exhibit both pluralism and consensus. It follows that we cannot make the inference to an upper boundary for $\bar{r}$. Thus, we cannot argue that G3's prediction holds. Again, for more informative answers, we would have

to simulate societies with branching.

This concludes my discussion of the research hypotheses L and G. Let's summarise the findings.

5.3 Summary

In this chapter, I have presented the results of the simulation study and tested the research hypotheses with respect to them.

In section 5.1, I have used so-called ternary heatmaps to present the arithmetic means of acceptance rates and the different measures for pluralism in the PC1-subsociety. These heatmaps lump together societies with the same numbers of connection types to PC1 into hexagon-shaped bins. The heat of each hexagon indicates the arithmetic mean of the respective metric. I could have chosen any PC to present the results (see appendix D), but PC1 appears in all structures, so considering it gives the most information. In what follows, I summarise the results as holding for an arbitrary PC.

I have described and explained how the number of connection types to PC influence consensus on PC and PC-pluralism in the respective societies. The general upshot has been that consensus on a PC is mainly facilitated by comprehensive doctrines that support PC. For pluralism in the PC-subsociety, on the other hand, there is a more complex picture. According to the two distribution-sensitive pluralism measures, entropy and strength of the weak, PC-pluralism is facilitated, first, by comprehensive doctrines that support PC. Second, it's facilitated by comprehensive doctrines that are neutral about PC, but only if there are no supportive doctrines in the structure. If exactly one doctrine supports PC, then there is little PC-pluralism. According to the distribution-insensitive measure option count, on the other hand, PC-pluralism is facilitated mainly by the absence of doctrines that are incompatible with PC. The numbers of neutral vs. supportive doctrines is not important. Finally, it's important to stress that these results are particularly preliminary for the linear model. Here the results are very sensitive to both society size and n_{PC}, meaning that we need more data to see whether the findings are robust. Despite this caveat, however, I think that these results about consensus and pluralism are by and large plausible and can be expected to give informative answers about the research question and

hypotheses.

In section 5.2, I have tried to show to what extent the research hypotheses L1–3 and G1–3 are falsified by or consistent with the present data. L1–3 state that potential *local* overlapping consensus in the different senses are improbable if it's not the case that most comprehensive doctrines in the structure support PC. G1–3 state that potential *global* overlapping consensus in the different senses are improbable if it's not the case that most comprehensive doctrines in the structure support PC. It turned out that for the purpose of testing these hypotheses, it is sensible to present the data in a slightly different way, namely, by plotting the ternaries with the median values and not the arithmetic mean values.

Given these ternaries, L1 is falsified by the data. There are hexagons in which a potential local overlapping consensus on PC in the weak sense is probable even though not a single doctrine supports PC. In particular, this holds for the hexagon with only neutral connections to PC. L2, on the other hand, cannot be falsified by the present data, because for each society only one tuple is drawn from the space of justified belief systems. This is not enough information to tell whether a potential local overlapping consensus in the strong sense is probable or not. However, we might be able to say something about L3, i.e. the hypothesis that it is improbable that a potential local overlapping consensus outside of the support area will be of high grade, i.e. of grade $r \geq 0.5$. However, for this we need what I called the indifference assumption: On average, LocalMRE does not favour tuples with high or low PC-pluralism but is indifferent regarding this property of tuples. If this holds, we can infer that the societies in the neutral corners have an average grade of $\bar{r} \geq 0.5$, which falsifies L3. If this assumption does not hold, however, things are less clear. We might need to simulate societies with branching.

Since consensus on PC is mainly facilitated by supportive doctrines, the present data cannot falsify the hypotheses G about potential global overlapping consensuses. What's more, however, we can even show that the predictions of some of these hypotheses are true in the present data. In particular, we can show that G2's prediction is true, i.e. if it's not the case that most doctrines support PC, then it is improbable that there is a potential overlapping consensus in the strong sense. Likewise, G3's prediction can

	weak sense	strong sense	grade $r \geq 0.5$
potential local OC	falsified	not enough data	falsified*
potential global OC	not enough data	verified	verified*

Table 5.1: This table summarises what the data says about the prediction that the different kinds of potential overlapping consensuses on PC are improbable if it's not the case that most comprehensive doctrines support PC. To say that a hypothesis is verified here only means that its prediction holds in the present data. Entries marked with (*) are subject to some variant of the indifference assumption about how LocalMRE samples spaces of justified belief systems.

be shown to hold, because outside of the support area, $\bar{r} < 0.5$. However, this is again subject to a variant of the indifference assumption: On average, LocalMRE is indifferent to whether tuples exhibit both PC-pluralism and consensus on PC. However, G1's prediction cannot be shown to be true for the same reason that L2's prediction cannot be shown to be false: There is not enough data. Thus, outside of the support area we know nothing about the probability of a potential global overlapping consensus in the weak sense.

Table 5.1 summarises these findings.

I have highlighted that studying various forms of the indifference assumption is crucial. Not only can the present data be further analysed. If LocalMRE's sampling is robust in the relevant way, then we can generally learn much about the (average) grades of societies simulated without branching, i.e. without calculating every possible fixpoint for every agent in every society. This enables us to set up future simulations studies in a frugal way.

The bottom line of the present chapter is that the data gives a mixed picture of the hypotheses: A potential local overlapping consensus in the different senses can be and sometimes is probable even if supportive doctrines are absent. The same thing cannot be said about the probability of a potential global overlapping consensus in the different senses. We have no reason to think that such overlapping consensus can be probable if supportive doctrines are absent.

This chapter was exclusively concerned with the dry, technical analysis of the study results. In the next chapter, after once more reviewing the main philosophical commitments of the present thesis, I will discuss what these results mean for political liberalism.

Chapter 6

Conclusion and outlook

This thesis began with the following plausible statements:

Pluralism	Societies are often pluralist, i.e. the citizens hold a diversity of worldviews.
Consensus	It would be good if citizens in a society agreed on constitutional essentials concerning the procedure and limits of political decision making.
Justification	It would be good if citizens in a society were justified in holding their beliefs.

I argued in the introduction that sometimes there is a tension between these statements. In particular, sometimes the pluralism of worldviews is such that it stands in the way of a consensus on constitutional essentials where all who agree are justified in holding their beliefs. If this is the case, then there are four ways of dealing with that: We can either do without consensus, or do without justification, or abolish pluralism altogether, or bring about conditions such that the pluralism does not stand in the way of consensus and justification. The first three options, I argued, are undesireable. The argument in short: If we give up the consensus condition, we risk societal instability. Abolishing or preventing a pluralism of worldviews requires oppressive means that are not available to liberal democracies, or it causes serious moral costs of other kinds. Forcing a consensus in a pluralist society by ignoring the justification conditions requires, again, oppressive means or leads to societal instability or both.

Thus, it is of great importance to investigate the fourth option by finding circumstances under which there can be pluralism, consensus, and justification. In John Rawls's terms, we need to find circumstances such that an *overlapping consensus* is possible. In this thesis, I have attempted to contribute to this goal. My focus has been on the epistemological aspect of the task: What needs to be the case such that the justification criterion permits for a constellation of belief systems to exhibit both pluralism and consensus? In what follows, I first give an overview of central points in the thesis, before drawing philosophical consequences and discussing next steps.

6.1 Overview

My methodology for investigating this question relies on formal epistemology. In particular, the thesis can be divided into two parts. In the first part, the *philosophical* part if you will, I developed a definition of the relevant kind of justification plus a handful of definitions for different stages of an overlapping consensus, and I formulated a general research question accompanied by some more specific research hypotheses. I did so mostly by discussing the Rawlsian account of overlapping consensus. In the second part, the *formal and computational* part if you will, I first gave formal explications of the notions of justification, consensus and pluralism, and put them together to yield formal explications of the different definitions of overlapping consensus. Then, I presented and discussed a simulation study that is designed to uncover the influence of the dialectical situations of the citizens on the possibility of an overlapping consensus. I will go through both parts of the thesis in some more detail, because doing so will give us an overview of all important assumptions upon which the results of the study rest.

Chapter 2 is about the philosophical foundations of the thesis, some of which I took from Rawls. The first and most important idea I drew from Rawls's political liberalism is the idea of an *overlapping consensus*: The pluralism in liberal democracies threatens their stability. The solution to this problem is that citizens, despite the differences in their worldviews, agree on a political conception of justice. This conception includes, most importantly, answers to the most basic questions of their constitution. It is a central part of the Rawlsian account of an overlapping consensus, and *the* central

assumption of this thesis, that citizens in an overlapping consensus are *justified* in accepting the shared political conception of justice. And they need not just any kind of justification: For example, the pragmatic justification of a compromise, or *modus vivendi* as Rawls calls it, is not sufficient, because its stability presupposes a power balance that may shift. Instead, for a greater stability of this consensus, citizens need to be morally justified in holding the political conception by integrating it into their system of moral beliefs as a whole. Rawls calls this 'full justification'. Even though I accept this solution to the problem of pluralism, I am not in any way committed to Rawls's particular ideas on how to characterise comprehensive doctrines (which is his term for what I called worldviews above) or political conceptions of justice. This is because in modelling the situation in the formal part of the thesis I adopt a purely structural perspective, thus, this model is compatible with many views on these matters. I take this to be a strong suit of the present research.

In addition to adopting the general Rawlsian idea of an overlapping consensus, I developed several relevant distinctions between different kinds of overlapping consensus. These distinctions can neither be found in Rawls's work nor in works of other philosophers, at least as far as I know. The first and perhaps most important distinction is that between an actual and a potential overlapping consensus. There is an actual overlapping consensus iff the citizens hold justified moral belief systems that agree on a political conception of justice whilst disagreeing on other moral matters, i.e. whilst exhibiting a pluralism of comprehensive doctrines. But suppose that citizens do not hold justified belief systems. Then one can still sensibly and with practical interest ask: Suppose for each citizen we know which belief system is justified for them. Would the resulting *combination* of justified belief systems form an overlapping consensus? In other words: If it *were* the case that each citizen holds the belief system that is justified for them, would there be a pluralism of doctrines and a consensus on a political conception? If the answer is yes, then there is a *potential* overlapping consensus. We might then try to bring about an actual overlapping consensus by bringing it about that each citizen adopts the belief system that is justified for them.

As I just phrased it, there is just one belief system that is justified for each agent. But there might be several. And, given the explication of justification

I proposed, there can indeed be several such belief systems. Thus, there is not just one constellation of justified belief systems in a society, but several. These possible combinations of justified belief systems form the *space of justified belief systems* of a society. Any combination is a point in this space and each such point is represented by a tuple of justified belief systems. Each position in the tuple corresponds to one citizen and can be filled with any belief system that is justified for that agent. As a consequence, there are different senses of a potential overlapping consensus, depending on how many of these tuples exhibit a pluralism of doctrines and a consensus on the political conception. If there is at least one such tuple, then there is potential overlapping consensus *in the weak sense*. If all tuples satisfy this condition, then there is a potential overlapping consensus *in the strong sense*. If a proportion $r \in [0,1]$ of the tuples satisfy the condition, then there is a potential overlapping consensus *of grade r*. These different senses correspond to different conditional probabilities. For example, if there is a potential overlapping consensus of grade r, then it is appropriate to say: If it were the case that all citizens hold a belief system that is justified for them (and we have no other relevant information), then with a probability of r there would an actual overlapping consensus.

The second distinction I drew is that between a global and a local overlapping consensus. Global here means 'society-wide' and is plainly what we are ultimately interested in. But even if there is no potential global overlapping consensus in whatever sense, it might still be sensible and of practical interest to ask: Is there a part of that society such that citizens (justifiedly) agree on the political conception whilst holding a pluralism of comprehensive doctrines? If there is, then in this society a justified consensus on the conception is in some sense *compatible* with a pluralism of doctrines. This is of practical interest, because we might try to turn this local overlapping consensus into a global overlapping consensus by, abstractly speaking, identifying the relevant circumstances that lead to the overlapping consensus in the respective part of society and try to bring it about that these circumstances hold on the rest of the society as well. Of course, this local kind of the potential overlapping consensus has different senses as well: There is a potential local overlapping consensus in the weak sense iff there is at least one tuple of justified belief systems such that among all citizens that agree on the political

conception there is a pluralism of comprehensive doctrines. There is one in the strong sense iff for all tuples there is a set of such citizens and there is one of grade r iff for a proportion $r \in [0, 1]$ of all tuples there is a set of such citizens.

Since I simulated *artificial* societies in the formal part of the thesis, I have here no use for the concept of an 'actual overlapping consensus', because there are no 'actual citizens' that do or do not 'actually' hold belief systems. Thus, the present thesis was focused on the remaining six kinds of overlapping consensus: the potential local overlapping consensus in the weak/strong/graded sense and the potential global overlapping consensus in the weak/strong/graded sense.

The second part of the philosophical foundations of this thesis was concerned with the relevant notion of justification, i.e. the notion of moral justification or justification of moral belief systems. I adopted an equilibrationist account of justification. In doing so, I followed Rawls himself, but also many other philosophers who are not directly concerned with political liberalism. Equilibrationism says, roughly, that moral beliefs are justified iff they are the result of applying the method of reflective equilibrium or can be reconstructed as such. In particular, the result of this method is supposed to be a *coherent* system of moral beliefs. Of course, these statements leave much to be clarified, so I committed to the following claims:

- **Dialectical Situation** The dialectical situation of an agent is the totality of views and arguments that the agent has to consider during equilibration such that the outcome can count as justified. A dialectical situation includes at least the views and arguments that are publicly debated in the agent's society, or so I argued. These form the *common core* of all citizens' dialectical situations.

- **Reconstructionism** Beliefs are justified iff they *could have been* the result of an equilibration process. (MRE is a test for whether beliefs are justified, no matter how they were generated.) One consequence of this is that the present thesis is concerned with propositional, not doxastic justification.

- **Epistemic consequentialism** The degree to which a belief system is in the state of reflective equilibrium is the feature that is deemed exclus-

ively epistemically valuable. The method of reflective equilibrium is simply a means to an end, namely increasing epistemic value.

- **Bounded Rationality** Epistemic agents are non-ideal. They have limited cognitive resources, etc. As a consequence, an agent's justification only requires that their belief system could have been the result of applying a *feasible and effective* method for increasing epistemic value.

These commitments gave us the following definition of (propositional) justification:

Definition 5 (Propositional Justification: equilibrationist, reconstructionist, consequentialist, non-ideal)**.** Let B be the set of all possible (moral) belief systems. Let a be an agent in dialectical situation D with initial commitments $C_0^a \in B$. Then the set $J \subset B$ of *belief systems propositionally justified for* a is defined as: $b \in J$ iff b could have been the result of applying a feasible and effective equilibration method starting from C_0^a and considering D.

Finally, I developed a research question, accompanied by a handful of research hypotheses. This research question and the accompanying hypotheses have guided both the formal explications of an overlapping consensus as well as the design of the simulation study. My general interest is to study the influence of the common core of the citizens' dialectical situations (i.e. the publicly debated views and arguments) on the possibility of a potential overlapping consensus. More precisely:

> **Research Question** Which kinds of inferential connections between the publicly debated comprehensive doctrines and a (publicly debated) political conception of justice make a potential overlapping consensus on this conception possible?

Public debate is important for democracies and can be expected to have a strong influence on the possibility of an overlapping consensus (due to its influence on the dialectical situations of the citizens). As a consequence, it is of great importance to discuss how to conduct public debate. One aspect of this is to investigate under which conditions public debate fosters an overlapping consensus.

I developed several research hypotheses that are important to test, or so I argued. It seems initially plausible that if the dialectical situations of the

citizens contain many comprehensive doctrines that *support* a conception, then an overlapping consensus should be possible if at all. Indeed, it seems that Rawls himself supposes that only comprehensive doctrines that support a conception can be part of an overlapping consensus on this conception. Formulated in terms of dialectical situations (and made more precise in various ways), this gave us the following hypotheses regarding the different kinds of overlapping consensus. Let the political conception PC be given and fixed.

> **Hypotheses L** If it's not the case that most comprehensive doctrines in the (common core of the) dialectical situations support PC, then
>
> 1. it is improbable that there is a potential local overlapping consensus on PC in the weak sense.
> 2. it is improbable that there is a potential local overlapping consensus on PC in the strong sense.
> 3. it is improbable that there is a potential local overlapping consensus on PC of grade $r \geq 0.5$.
>
> **Hypotheses G** If it's not the case that most comprehensive doctrines in the (common core of the) dialectical situations support PC, then
>
> 1. it is improbable that there is a potential global overlapping consensus on PC in the weak sense.
> 2. it is improbable that there is a potential global overlapping consensus on PC in the strong sense.
> 3. it is improbable that there is a potential global overlapping consensus on PC of grade $r \geq 0.5$.

I have argued that these hypotheses, if true, pose high standards for a public debate that fosters an overlapping consensus. In particular, if we want to conduct public debate such that it fosters an overlapping consensus on a given political conception, then we would have to exclude both doctrines that are incompatible with the conception and doctrines that are neutral

about it. This exclusion, especially the exclusion of merely neutral doctrines, seems to go strongly against the idea of a public debate open to all. Thus, it is important to test these hypotheses.

Chapter 2 was devoted to giving explications of the different notions of a potential overlapping consensus. In particular, I gave an explication of the notion of justification that respects the philosophical commitments from the previous chapter. For doing so I relied on the formal model of reflective equilibrium put forth by Beisbart et al. (2021), which I call 'BBB model'. The model is based on the theory of dialectical structures by Betz (2021). Accordingly, a dialectical structure (a set of sentences connected by arguments) is the background for any equilibration process. The belief system of an agent is represented by a pair of positions in this structure: The commitments and the theory of the agent. This pair of positions, called the epistemic state of the agent, can have different degrees of equilibrium, i.e. different degrees of being in the state of reflective equilibrium. This notion of degrees of equilibrium is explicated by the so-called *achievement function*. The achievement function reflects in how far the commitments are derivable from the theory, how systematic the theory is, and whether there is a sufficient tie to the initial commitments of the agent. There are two versions of the achievement function: a *quadratic* and a *linear* version, depending in the inner workings of the function (see appendix A for more). An algorithm (i.e. a method of reflective equilibrium) is given that indicates how to maximise the achievement function, namely, via a process of mutual adjustments of theory and commitments to each other. This process is a form of semi-global optimisation: When the theory is adjusted, then we consider *all* logically possible theories, likewise for the commitments.

I have suggested to change this algorithm for the purposes of this thesis. In particular, it does not fit the bounded rationality perspective I am embracing, because it requires significant computational resources. Instead, I have argued that a form of local optimisation is better suited for investigating the research question. This adapted algorithm does not consider all possible theories, but only those in the close neighbourhood of the current one, likewise for the commitments. This local algorithm is my explication of the notion of a 'feasible and effective' method for increasing coherence. Using this adapted model, I have given an explication of the notion of a justified belief system:

Roughly, the set of justified belief systems for an agent is the set of possible outcomes of applying the local algorithm to the agent's initial commitments. It should be noted however, that this explication is, strictly speaking, *two* explications, because there are two versions of the achievement function that the algorithm may use: the quadratic and the linear one.

Given this explication of justification, all that was left to do was to explicate the notions of consensus and pluralism. I gave a measure for the consensus (on a given political conception *PC*) of a tuple of belief systems, and three measures for the pluralism of comprehensive doctrines among the belief systems accepting a given conception, i.e. three measures of *PC*-pluralism. The consensus measure, called *acceptance rate*, is very simple: It returns the number of belief systems accepting *PC* divided by the total number of belief systems. (It's also normalised to range from 0 to 100.) The three pluralism measures, option count, strength of the weak, and entropy, each focus on a different aspect of the notion. *Option count* is a very minimal notion, roughly, it counts the number of doctrines that are accepted in at least one belief system. I highlighted that option count is in a sense *distribution-insensitive*, because it does not mind whether a doctrine is accepted once or a hundred times. The thought behind *strength of the weak* is that dominance is the enemy of pluralism. Thus, roughly, it measures how many fixpoints do *not* accept the strongest comprehensive doctrine. Finally, *entropy* is a tool from information theory that can be taken to measure how homogeneous a distribution is. Here, we are measuring how homogeneous the distribution of agents over the doctrines are. Entropy is maximal iff all doctrines are accepted exactly the same amount of times. Since this holds also for parts of the distribution while holding the rest fixed (a feature called additivity), this is a generalisation of the idea that dominance is the enemy of pluralism, or so I argued. Strength of the weak and entropy are, in contrast to option count, distribution-sensitive.

Of course, the three pluralism measures work differently and I did not pick one as the most important one. Instead, I used all three of them. Together with the measure for consensus and the explication of justification, I gave explications of the different notions of a potential overlapping consensus. Since there are two kinds (global/local) and three senses (weak/strong/graded) of a potential overlapping consensus, as well as two

achievement functions (quadratic/linear) and three pluralism measures (option count/strength of the weak/entropy), this yielded a total of 2×3×2×3 = 36 explications of the notion of a potential overlapping consensus.

With these tools in hand, we were in a position to design a simulation study and analyse its results. In chapter 4, I presented the design. There are two parts to this design: First, I characterised the kinds of societies to simulate. Second, I detailed how to sample the resulting possibility space such that I am in a position to address the research question and test the hypotheses. Regarding the first point, any society consists of a dialectical structure (the common core of the citizens' dialectical situations) and a bunch of initial commitments (the citizens). This is a central idealisation of the present study: I assumed that the common core of the citizens' dialectical situations is all there is to these situations and, as a consequence, all citizens share the same dialectical structure. It remains to be seen whether the results are robust when the study design is de-idealised. This was not the only idealisation. In particular, I focused on unrealistically small toy structures that resemble the important features of the more realistic big counterparts. In particular, this resemblance consists in these structures containing a bunch of comprehensive doctrines and some political conceptions of justice and different connections between these two kinds of sentences. For each pair of comprehensive doctrine and political conception, there are three possible connections: Either the doctrine supports the conception, or is incompatible with it, or is neutral about it/there is no connection. These connections between doctrines and conceptions I called the structure's *head*. The structure's body comprises the inferential connections between the doctrines and other sentences in the structure as well as the conceptions and other sentences. In the end, I gave a definite list of conditions for possible societies in my study.

Regarding the second point, sampling the resulting space of possible societies, my focus was on the structures' heads, because the reseach question and hypotheses are precisely about the influence that the structure's head has on the possibility of an overlapping consensus. Unfortunately, there are too many possible heads to simulate all of them, but using a combinatorial "trick", I was able to reduce the number of relevant heads drastically. From the remaining number of combinations, I drew a random sample. In order to

isolate the influence of a particular head on the possibility of an overlapping consensus, 50 random bodies are drawn and for each resulting structure, 30 initial commitments are drawn. For each agent, i.e. each of the 30 sets of initial commitments, the locally optimising algorithm was applied once with quadratic and once with linear achievement function. Thus, for each model and agent, only one of all possible fixpoints was calculated. That is, for each model and society, only one tuple was drawn from the space of justified belief systems. For each tuple, consensus and pluralism was calculated using the different measures I introduced.

In chapter 5 I would then average the results for consensus and pluralism in order to average out the influence of bodies and initial commitments. So let's turn to these results. For the global kinds of overlapping consensus on a political conception *PC* we need both consensus on *PC* and pluralism amongst the fixpoints accepting *PC*, i.e. *PC*-pluralism. For the local kinds, we only need *PC*-pluralism.

When reviewing the arithmetic means of the acceptance rates, we have seen that consensus correlates mainly with the number of doctrines in the dialectical situation that *support PC*. A direct consequence of this is that the global variants of the research hypotheses, i.e. G1–3, cannot be falsified by the data, because they precisely claim that a potential global overlapping consensus in the different senses is improbable if it's not the case that most doctrines support the conception. (To make this precise, we also had a look at the median values.) What's more, we can also say that G2–3's predictions turned out to hold in the data. That is, given that we know nothing of a society but the numbers of doctrines that support *PC*, are neutral about it, and incompatible with it, it is improbable that there is a potential global overlapping consensus on PC in the strong sense or of high grade. (The point about potential global overlapping consensus of high grade required a certain assumption, see below.) If this result turns out to be robust, then it sets a high standard for an overlapping consensus: A potential global overlapping consensus is only probable if most publicly debated comprehensive doctrines support *PC*.

When reviewing the arithmetic means for *PC*-pluralism, we have found that there is a significant difference between the pluralism measures, i.e. the different ways to explicate the notion of pluralism. According to the

distribution-insensitive measure, option count, *PC*-pluralism is comparatively easy to come by. It mainly correlates with the *absence* of doctrines (in the dialectical situation) that are *incompatible* with *PC*. Thus, if all or most doctrines are neutral about or supportive of *PC*, then there will likely be a high degree of *PC*-pluralism. According to the distribution-sensitive measures, entropy and strength of the weak, the same holds, but with an additional requirement. The requirement is that it is not the case that exactly one doctrine supports *PC* while the others are neutral about it or incompatible with it. In this case, there will be little *PC*-pluralism. But if either many doctrines are neutral or many doctrines are supportive, then there will be *PC*-pluralism. (Many incompatible doctrines are always bad.) As a consequence of these findings for the different measures, we can infer that some of the local variants of the research hypotheses are falsified by the data. (Again, to make this precise, we also had a look at the median values.) In particular, we can say that a potential local overlapping consensus *in the weak sense* and *of high grade* are probable even if all doctrines are neutral about *PC*, i.e. even if it is not the case that most doctrines support *PC*. However, the claim about an overlapping consensus of high grade (just like its global counterpart, see above) depends on a non-trivial *indifference assumption* about how the algorithm samples spaces of justified belief systems. This assumption is as of now unverified, but I argued that it is worthwhile to test it. If it turns out to hold, then we may set up future simulation studies in a computationally frugal way.

On the bottom line, the results give a mixed picture: The standards for a potential *global* overlapping consensus on a political conception PC are rather high, requiring most publicly debated doctrines to support PC. The standards for a potential *local* overlapping consensus are lower, only requiring (roughly) that most doctrines are not incompatible with PC. Of course, we are ultimately after a global, not local overlapping consensus. In the next section, I discuss why the findings nonetheless give us hope that an overlapping consensus is a viable option.

6.2 Philosophical interpretation

This concludes my overview of the present thesis. Before moving on to potential follow-up studies, I wish to attempt to distil something like a takeaway message of this thesis. All questions of robustness, idealisations, unverified assumptions, etc, aside: Does the present thesis give us hope that an overlapping consensus is a viable possibility? I think it does. In particular, it shows that there are conditions such that a potential global overlapping consensus is likely. But it also shows that there are significant challenges to realising an overlapping consensus, because these conditions set a high standard. In what follows, I want to make suggestions concerning what consequences would have to be drawn if the present results turned out to be robust. These suggestions are speculative. That is, they should not be taken as recommendations that are based on robust scientific results. Instead, they are supposed to highlight the importance of and a potential direction for further research. Importantly, when discussing the results concerning potential *local* overlapping consensus, it will turn out that the high standard for potential global overlapping consensus might not be necessary after all. This further bolsters hope that an overlapping consensus is a realistic option.

Potential global overlapping consensus

The general gist of the results is this: *If* most comprehensive doctrines in the common core of citizens' dialectical situations, i.e. most publicly debated doctrines, are supportive of a given political conception, *then* things look good. A vast majority of the simulated societies exhibits a high degree of consensus. A vast majority also exhibits a high degree of *PC*-pluralism. This means that a potential global overlapping consensus is likely. I have not actually given numbers for this probability, because given the data presented in the previous chapter 5 we do not know how many of the simulated societies exhibit *both* a high (enough) consensus *and* a high (enough) *PC*-pluralism. Nonetheless, we can expect this probability to be relatively high.

However, if it is *not* the case that most publicly debated comprehensive doctrines are supportive of a given political conception, i.e. most of them are neutral about or incompatible with it, then things look badly. As I have argued at length, the present results suggest that a potential global

overlapping consensus (in the strong sense and of high grade) is unlikely under these circumstances. This is worrisome. Let's suppose that this result turns out to be robust (though below I sketch a way to challenge it). How could a society deal with that?

Of course, if it is in fact the case that most doctrines in public debate are supportive of the implemented political conception of justice, then all is good for the time being. But if this is not the case, or public debate begins to change such that it is not the case, then what? Roughly, there are two options. We can either try to influence public debate, thereby changing citizens' dialectical situations, or influence citizens' initial commitments.

1. We can influence the dialectical situations, either by

   (a) excluding (some) incompatible or neutral doctrines from public debate. As a result, there are less such doctrines in the dialectical situations of the citizens and, *relatively speaking*, more supportive doctrines. This increases the probability that the initial commitments of any given citizens is matched best by a supportive comprehensive doctrine and, as a result, this supportive doctrine (together with the supported political conception) appears in the belief system(s) justified for them. Obviously, excluding incompatible and neutral doctrines from public debate is highly problematic, because it is illiberal. As I have already mentioned in section 2.2.6, public debate should be open to all views and arguments (paradigmatically: Habermas, 1990, p. 89). Any constraints are in dire need of justification. For incompatible doctrines, we might be able to argue that their exclusion is warranted, because the cost of them destabilising society is too high and perhaps one can even argue that some of these views are 'objectively' wrong and public debate does not profit from their inclusion. (Popper famously argued along these lines when treating the *paradox of tolerance* (2012, p. 581). Rawls makes a similar, though more cautious remark (TJ 220). For an opposing view see Walzer (1997, pp. 80f).) I think that to some extent, this exclusion of incompatible doctrines is already realised in liberal democracies. For example, uttering fascist views is often considered a no-go in talkshows,

parliaments, etc. But for neutral doctrines, this is harder to argue. In particular, there seems to be nothing wrong with, e.g., a purely spiritual worldview that is mostly concerned with the non-public sphere and not with politics. Excluding such views might seem *too* illiberal. Of course, it is not so clear how the relevant trade-offs are to be made. But it would be best if we didn't have to make such trade-offs.

(b) including more supportive doctrines in public debate, thereby giving citizens more supportive options that might fit their initial commitments better than the neutral or incompatible ones. As a result there are again, relatively speaking, more supportive doctrines in citizens' dialectical situations. This is, of course, the preferable option, because including views and arguments is not in itself illiberal. But supportive doctrines don't grow on trees. Perhaps this is a task for moral philosophers: Devise comprehensive moral doctrines supportive of a liberal democratic constitution that are an appealing alternative for citizens with initial commitments that also fit well with comprehensive doctrines that are neutral or incompatible with such a constitution.

2. If there are *some* supportive doctrines in the common core of citizens' dialectical situation, then we can try to influence citizens' initial commitments such that they fit best with these supportive doctrines instead of neutral or incompatible ones. (How to do that will obviously in part depend on what exactly initial commitments are, see section 2.2.1. Are they considered judgments (Rawls), initially tenable commitments (Elgin) or intuitions (Lewis)?) However, it seems difficult to do that without relying on propaganda, indoctrination or oppressive means. Freedom of opinion, religion, etc. are high goods in liberal democracies. Most direct forms of interference from an institutional level will quickly and perhaps rightfully incur the accusation of being illiberal or oppressive. A legitimate form of interference might be *deradicalisation programmes*. These offer support to citizens who want to leave extremist groups propagating incompatible doctrines, e.g. neo-Nazi or Islamist groups. Once someone has left their extremist group, they

may be more free to change in a way such that their initial commitments fit best with a supportive doctrine instead of an incompatible one. And indeed, many adopt a more modest worldview afterwards, as the growing and encouraging body of research on the theory and effectiveness of deradicalisation shows (Bjørgo and Horgan, 2009; Rabasa et al., 2010; Koehler, 2017). Importantly, if these programmes are offered and not mandated, then they are not illiberal or oppressive. However, they only address incompatible and not neutral doctrines. Also, it is not clear how cost-effective they are. Again, it would be best if there are other or additional solutions.

This list of options is likely not comprehensive and, again, it is not clear what the relevant trade-offs would be. But the problems associated with these options show that it would be better if we could show that the result about potential global overlapping consensus, i.e. that incompatible and neutral doctrines make it unlikely, is *not* robust. Luckily, an investigation of the results for a potential *local* overlapping consensus might give us an idea of how to avoid these problems.

Potential local overlapping consensus

First of all, if most publicly debated doctrines are *incompatible* with the given political conception, then *PC*-pluralism in the simulated societies is very low, i.e. a potential local overlapping consensus in the strong sense or of high grade is very unlikely. Strictly speaking, I have not argued for this in the last chapter (since I was exclusively concerned with the research hypotheses about supportive doctrines), but the argument is straightforward.

This is, again, not good (though perhaps not that surprising). As a consequence, it seems we will not get around dealing with the problem of incompatible doctrines in public debate (if there are such doctrines), see the options described above. But what about neutral doctrines? The results show, as I have argued at length, that a potential local overlapping consensus at least in the weak sense, if not of high grade, is probable even if most or all comprehensive doctrines are neutral. Thus, even though the results suggest that there is a high standard for a potential *global* overlapping consensus (most publicly debated doctrines must be supportive), the standards for

a potential *local* overlapping consensus are lower (most publicly debated doctrines must be supportive *or* neutral).

But what does this finding about the lower standards for a potential local overlapping consensus help us if we are ultimately interested in a *global* overlapping consensus and such a consensus requires the higher standards? Remember what I said about the practical relevance of a potential local overlapping consensus: We might try to turn this local overlapping consensus into a global overlapping consensus by, abstractly speaking, identifying the relevant circumstances that lead to the local overlapping consensus in the respective part of society and try to bring it about that these circumstances hold in the rest of the society as well. Put differently, there might be favorable conditions such that a potential global overlapping consensus is likely even if many publicly debated doctrines are neutral. (Of course, this would mean that in the simulated societies conditions are not favorable.)

I have a hypothesis about what these favorable conditions could be. Consider a society with (publicly debated) comprehensive doctrines that are all *neutral* about the political conception of justice. As we have seen, my study suggests that in this case a global overlapping consensus is improbable, but a local overlapping consensus is probable. I suspect that one thing that the citizens participating in a local overlapping consensus have in common is that they have initial commitments that fit the political conception well enough such that it is worthwhile, in terms of coherence, to accept the conception as a part of their theory, alongside some neutral doctrine. (At least, this seems to hold for the quadratic achievement function, see section 5.1.2.) As a consequence, we might try to test whether a potential *global* overlapping consensus is probable if *all* (or most) citizens have initial commitments that fit the political conception well. In the present study, this is not the case, because the initial commitments were randomly generated with homogeneous probability distribution. Thus, even though in the present study a global overlapping consensus requires that all doctrines support the conception, it might turn out that if many citizens have initial commitments that fit the conception, then a global overlapping consensus does *not* require this high standard and instead, it suffices that all doctrines support the conception or are neutral about it. Note that this concerns only the initial commitments regarding constitutional essentials (i.e. the 'PPS' in the study design) and not

the ones concerned with the non-political (i.e. the 'PS' in the study design).

In the next section 6.3 I suggest a follow-up study that tests this hypothesis. For now, let's suppose that this is indeed the case. Then if these favorable conditions are met, all is good for the time being. If not, then we can ask ourselves how to bring about these favorable conditions without relying on propaganda, indoctrination or other oppressive means which are not available to a liberal democracy. Again, this will depend on what you take initial commitments to be and also on what is legitimate for a liberal democratic society to do. In the end, these will be difficult questions that I cannot answer here. But here are a few ideas:

1. Civic education:

   - Inform citizens about the constitutional essentials (or enable them to inform themselves with reliable sources). How do elections work? How do the governing bodies form on different levels, i.e. the federal, state and municipal level (where applicable)? How is the separation of powers institutionalised? What is the standing of human rights in the constitution? Only on the basis of such knowledge can citizens form appropriate initial commitments.
   - Inform citizens about life under alternative, particularly authoritarian, forms of government (or enable them to inform themselves with reliable sources). This information can be both about currently existing societies (e.g. China or Noth Korea) or historical societies (e.g. Nazi Germany). Highlighting the alternatives helps citizens appreciate the advantages of a liberal democratic regime.

   For further reading on civic and democratic education, see Culp et al. (2023), Peterson et al. (2020) and Macleod and Tappolet (2019).

2. Democratic participation: Enable and encourage citizens to participate in democratic structures. This includes participation in elections (exercise of active electoral rights), parliaments on all levels and city councils (exercise of passive electoral rights), involvement in political parties, but perhaps also to some extent companies (e.g. in the works council), administration of schools, universities, etc. The same goes for engagement in political activism (demonstrations, non-governmental or-

ganisations) and public debate. Democratic participation might foster citizens' self-perception as political agents who have an influence on their society and the rules that govern it. Perhaps this can be understood as participating in what Rawls called the *public political culture* in a society (see section 2.2.3). For further reading on political participation and education through activism, respectively, see Giugni and Grasso (2022) and Hall et al. (2012).

3. Improving the system: Of course, if the current political system has serious flaws, then this will make it more likely that citizens will blame the political conception of justice for these flaws (rightfully or not). Thus, making sure that 'everything works well' will foster trust or reliance in the political system (Budnik, 2018). This includes effective governance, well-functioning elections, but perhaps also fighting fake news and populism.

Let's suppose that some of these ideas (or further ones not listed here) are in fact effective at fostering initial commitments that are in alignment with the political conception of justice that is implemented in society whilst *not* being illiberal or oppressive in any way. Then we have a choice between *two* options: Either we try to bring it about that all publicly debated doctrines support the conception (in which case the initial commitments don't matter so much), or we try to bring it about that all publicly debated doctrines are at least compatible with (if not supportive of) the conception *and* most citizens' initial commitments fit the conception well. Compared to the present results, where the only option is that most doctrines support the conception, this would be an advance. But again, these considerations rest on the speculation that this condition (i.e. that all or most citizens' initial commitments align with the political conception of justice) is indeed favorable in the sense that a potential global overlapping consensus is likely even if most publicly debated doctrines are neutral.

Of course, the ideas for fostering a stably functioning liberal democracy presented in this section are not new. But the present research makes two key contributions. First, it clarifies what the *epistemic* roles of these ideas are with respect to an overlapping consensus. For example, excluding incompatible doctrines from public debate serves to exclude them from the dialectical

situations of the citizens such that citizens are not required to consider such doctrines in order to hold justified belief systems. This makes the justified acceptance of such doctrines less likely. Civic education and democratic participation serve to foster initial commitments that align with the respective political conception such that it is worthwhile, epistemically speaking, to accept this conception even if one's comprehensive doctrine is neutral about it. Thus, the present research highlights that these measures may play an important role not only for the opinion dynamics of consensus formation, but also for this consensus to be justified. Second, if the considerations above are correct, then this line of research contributes to informing relevant trade-offs. For example, as I have argued, it might not be necessary to exclude all non-supportive doctrines from public debate in order to safeguard societal stability. If we can show that it is sufficient to foster initial commitments that are in line with the political conception, then we can refrain from relying on such potentially illiberal measures, at least when considering the normative side of things.

I want to highlight, however, that in order to truly judge the viability of these solutions, we also have to have a look at the empirical side of things. In particular, we have to see what boundary conditions are given by the psychology or actual opinion dynamics of the citizens. This topic is outside of the scope of this thesis. Nonetheless, I want to stress that we need both normative and empirical research on overlapping consensus. If we rely exclusively on normative research, then we will never know how realisable an overlapping consensus really is given psychological and other boundary conditions. If, on the other hand, we rely exclusively on empirical, descriptive research, then we learn a lot about pluralism and consensus formation, but not necessarily about overlapping consensus since this is an inherently normative concept (via the justification condition). Thus, we need both normative and empirical research on overlapping consensus. In a sense, the present research, though not itself empirical, can be the basis for empirical research on the matter. In particular, it provides a clarification of the relevant concepts, including the normative concepts, in the form of different explications of the notion of an overlapping consensus. These explications, due to their formal precision, might be rather straightforwardly operationalised such that they are applicable in empirical research. Moreover, the present

line of normative research can guide the direction of empirical research by highlighting potential trade-offs that arise from the normative perspective (for example, the one described above). We can and should then add the empirical perspective on these trade-offs in order to see which solutions are most viable when both normative and empirical boundary conditions are considered.

But there is much work to be done before we can employ this strategy, so let's have a look at potential follow-up studies.

6.3 Follow-up studies

As I have stressed many times throughout this thesis, the results of the present study presuppose a myriad of assumptions. Some of them are philosophical assumptions, like the equilibrationist account of justification, some are modelling assumptions, like the theory of dialectical structures as a representation for beliefs, and some are idealising assumptions to keep things simple, like the shared dialectical situation of citizens in a society. In effect, there is a huge space of alternative ways to study the present research question and this space is filled with fog: We cannot see which answers these alternative ways would give. Given the assumptions of the present thesis, I have tried to clear the fog in one small area of this huge space.

But of course, we cannot with confidence give an answer to the general research question, if we have not cleared the fog in more than just one small area in order to see whether we always get the same or similar results. We need to check the *robustness* of the results. In this section, I lay out which robustness studies seem most interesting to me. There are at least three stages to checking robustness: Varying the study design, varying the model of reflective equilibrium, and considering completely different ways of explicating justification.

Vary the study design

This is the most natural next step and I have a ton of ideas for it. First and foremost, however, it should be investigated whether the so-called indifference assumption holds (see chapter 5). That is, it should be tested whether

LocalMRE is in fact not systematically sensitive to whether a tuple exhibits consensus or pluralism. If this assumption holds, then this will bolster the claims I made about potential local and global overlapping consensus of high grade. Also it will allow setting up future studies in a computationally frugal way. Arguably, testing the indifference assumption requires running some simulations with branching, i.e. for every random choice, follow up on all outcomes. It's not clear in how far this is technologically possible as of now and, if not, whether there is an alternative way to test the assumption.

Additionally, here are some ideas to vary the study design, though not necessarily in their order of importance:

- Drop some of the idealising assumptions for the dialectical structures, for example:
  - Allow for more complex arguments, i.e. arguments with more than one premise.
  - Let general statements, i.e. the doctrines and conceptions, be part of the initial commitments.
  - Allow for more classes of statements. In particular, allow for mid-level principles between doctrines and particular statements.
  - Allow for differences in the dialectical structures of different agents in the same society. Their belonging to the same society is then modelled by the common core of their dialectical situations.
- Consider the influence of the initial commitments. As of now, they are generated with homogeneous probability distribution. Citizens in real societies do not, of course, have such completely random initial commitments.
  - Model the influence of a public political culture. Living together in such a culture might lead to the citizens' initial commitments regarding the PPSs being similar to each other and close to the content of the political conception that is implemented in their society. This might significantly boost acceptance rates for this conception even when all doctrines are neutral about it, i.e. this robustness study tests the speculative hypothesis from the last section.

  - In general, try to see how sensitive the results are to inhomogeneities in the probability distribution of the initial commitments. For example, if few citizens have initial commitments that are close to the content of incompatible doctrines, then their presence in the structure might not impinge as much on consensus and pluralism (see suggestion 2 of dealing with the present results regarding potential global overlapping consensus).

- Investigate how the setup other than structures and initial commitments influences the results.

  - Simulate large societies with $n_{PC} > 1$. We can expect a different behaviour of the linear model here. As of now, due to limitations on computational power, I have only simulated large societies with $n_{PC} = 1,\ n_{CD} = 4$.
  - Slightly vary $\alpha_A, \alpha_S, \alpha_F$, i.e. the weights for account, systematicity and faithfulness in the achievement function. The current weights yield plausible results, but of course their exact values are somewhat arbitrary. We would not want to see a completely different behaviour when slightly varying them.

Vary model of MRE

Varying the model MRE is an important task, since the current formal model of reflective equilibrium, even though it is thoroughly tested, is just one plausible way of modeling MRE. We have little knowledge of how variants of this model work and whether they are more plausible. In any case, it is safe to say that this is not the only plausible variant of the model. Thus, we should have a look at how sensitive the results are to plausible changes of the model, for example:

- Add weights to the commitments of the agents.
- Explore algorithms other than the local algorithm I defined in chapter 3.3 that are plausible and seem to have a good balance of feasibility and effectiveness. A simple starting point would be to consider alternatives to $T_0 := \emptyset$ (e.g. Flick, 2022).

- Consider modifications and extensions to the achievement function. In particular, a first step could be to alter the measure for systematicity which can be greatly improved, I think, though this may require a recalibration of the $\alpha_A, \alpha_S, \alpha_F$. For some suggestions, see (Freivogel and Cacean, 2023, appendix C).

This is, of course, of interest not only to political liberalists, but also to theorists of reflective equilibrium.

Alternative explications of justification

In the long run I think it is important to also consider alternative frameworks for justification. In particular:

- Consider alternative models for reflective equilibrium:
  - Baumgaertner and Lassiter (2023) have put forth a basic model of reflective equilibrium.
  - Freivogel (2021) presents a formal model of MRE based on the AGM formalism.
  - Dellsén (2024) tries to cash out reflective equilibrium (partially) in terms of probabilities.
  - One might try to build a model for reflective equilibrium by utilising the formal accounts of coherence by Tersman (1993); Thagard (2000).
- Consider alternative explications for justification or rationality, i.e. explications that do not rest on the equilibrationist account of justification:
  - Degrees of belief: Bayesianism (Lin, 2024), ranking theory (Spohn, 2012), Dempster-Shafer theory (Dempster, 1968), etc.
  - Categorical belief: AGM belief revision theory (Huber, 2013), non-monotonic logic (Strasser and Antonelli, 2024), etc.

  For an excellent overview of the options, see (Genin and Huber, 2022).

This concludes my presentation of ideas for robustness studies. As a general rule of thumb, the further away a robustness study is from the original one, the more impressive are similar results. Suppose we conduct five robustness studies, varying the study design as the first step suggests. Suppose we get similar results, i.e. they turn out to be robust. That's good, it raises our confidence in the findings. But suppose, on the other hand, that we conduct five robustness studies that rely on completely different assumptions, e.g. by doing a Bayesian study, and AGM study, etc. Now, if these yield similar results, then that's much more impressive and it more so raises our confidence in the findings. The downside is, of course, that it is much more work.

You will see by now that the present thesis proposes a whole new research programme: Start robustness studies for the present results, try to think of new research questions and hypotheses concerning the possibility of overlapping consensus, connect them with the practical and empirical dimensions to check that we are on the right track, etc. As the questions and study designs evolve, we better understand how the employed models work and how overlapping consensus works. Piece by piece we uncover conditions that make an overlapping consensus more likely in real-world scenarios. This, in turn, contributes to fostering stability in liberal democracies without relying on coercive means.

The present thesis developed a conceptual foundation for this programme and produced first results that inspire further research.

Bibliography

Aczél, J., Forte, B., and Ng, C. T. (1974). Why the shannon and hartley entropies are 'natural'. *Advances in Applied Probability*, 6(1):131–146.

Ahlstrom-Vij, H. K. and Dunn, J. (2018). Introduction. In Ahlstrom-Vij, H. K. and Dunn, J., editors, *Epistemic Consequentialism*, pages 1–20. Oxford University Press, Oxford, UK.

Al Jazeera (2022). Iran denies Mahsa Amini, woman who died in custody, was beaten. https://www.aljazeera.com/news/2022/9/19/irans-police-denies-women-who-died-in-custody-was-beaten [Last accessed: 2023-12-13].

Alcalde-Unzu, J. and Vorsatz, M. (2011). Measuring consensus: Concepts, comparisons, and properties. *Studies in Fuzziness and Soft Computing*, 267:195–211.

AP Press (2022). Explainer: What kept Iran protests going after first spark? https://apnews.com/article/iran-protests-morality-police-explainer-b53475eda867a4158ac5032fe1b3e62e [Last accessed: 2023-12-13].

Bächtinger, A., Dryzek, J. S., Mansbridge, J., and Warren, M. (2018). Deliberative democracy: An introduction. In Bächtinger, A., Dryzek, J. S., Mansbridge, J., and Warren, M., editors, *The Oxford Handbook of Deliberative Democracy*, pages 1–32. Oxford University Press, Oxford, UK.

Baumberger, C. and Brun, G. (2021). Reflective equilibrium and understanding. *Synthese*, 198:7923–7947.

Baumgaertner, B. and Lassiter, C. (2023). Convergence and shared reflective equilibrium. *Ergo: An Open Access Journal of Philosophy*, 10(n/a).

Bayles, M. D. (1986). Mid-level principles and justification. *American Society for Political and Legal Philosophy*, 28:49–67.

Beauchamp, T. L. and Childress, J. F. (2013). *Principles of Biomedical Ethics*. Oxford University Press, Oxford, UK, 7th edition.

Beisbart, C., Betz, G., and Brun, G. (2021). Making reflective equlibrium precise: A formal model. *Ergo*, 8.

Beisbart, C. and Brun, G. (2024). Is there a defensible conception of reflective equilibrium? *Synthese*, 203(79):1–27.

Betz, G. (2014). *Debate Dynamics: How Controversy Improves our Beliefs*. Springer, Luxembourg.

Betz, G. (2021). *Theorie dialektischer Strukturen*. Klostermann, Frankfurt, Germany.

Bjørgo, T. and Horgan, J. G. (2009). *Leaving Terrorism Behind: Individual and Collective Disengagement*. Routledge, Abingdon, UK.

Boettcher, J. (2020). Deliberative democracy, diversity, and restraint. *Res Publica*, 26:215–235.

BonJour, L. (1985). *The Structure of Empirical Knowledge*. Harvard University Press, Cambridge, USA.

Bourget, D. and Chalmers, D. (2023). Philosophers on philosophy: The 2020 philpapers survey. *Philosophers' Imprint*, 0(0):1–53.

Bovens, L. and Hartmann, S. (2003). *Bayesian Epistemology*. Clarendon, Oxford, UK.

BPB (2021). Einstellungen zu Demokratie und Sozialstaat. https://www.bpb.de/kurz-knapp/zahlen-und-fakten/datenreport-2021/politische-und-gesellschaftliche-partizipation/330219/einstellungen-zu-demokratie-und-sozialstaat/ [Last accessed: 2021-12-31].

Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Sack, G., Fisher, S., Flocken, C., and Holman, B. (2017). Understanding polarisation. *Philosophy of Science*, 84(1):115–159.

Briesen, J. (2017). Evidentielle Einzigkeit in klassischer und formaler Erkenntnistheorie. *Zeitschrift für philosophische Forschung*, 71(2):183–222.

Budnik, C. (2018). Trust, reliance, and democracy. *International Journal of Philosophical Studies*, 26(2):221–239.

Carnap, R. (1963). *Logical Foundations of Probability*. University of Chicago Press, Chicago, USA, 2nd edition.

Carter, J. A. and Gordon, E. (2014). Objectual understanding and the value problem. *American Philosophical Quarterly*, 51(1):1–13.

CdV (2016). Diversity in germany: Study marking ten years of the diversity charter. `https://www.charta-der-vielfalt.de/fileadmin/user_upload/Studien_Publikationen_Charta/Diversity_in_Germany_2016_en.pdf` [Last accessed: 2024-06-19].

Christensen, D. (2007). Epistemology of disagreement: The good news. *Philosophical Review*, 116(2):187–217.

CNN (2021). Timeline of the coup: How Trump tried to weaponize the justice department to overturn the 2020 election. https://edition.cnn.com/2021/11/05/politics/january-6-timeline-trump-coup/index.html [Last accessed: 2023-12-13].

Coady, C. A. J. (1992). *Testimony: A Philosophical Study*. Clarendon, Oxford, UK.

Culp, J., Drerup, J., and Yacek, D., editors (2023). *The Cambridge Handbook of Democratic Education*. Cambridge University Press, Cambridge, UK.

Daniels, N. (1979). Wide reflective equilibrium and theory acceptance in ethics. *The Journal of Philosophy*, 76(5):256–282.

Daniels, N. (1996). *Justice and Justification. Reflective Equilibrium in Theory and Practice*. Cambridge University Press, Cambridge, UK.

Dellsén, F. (2024). Probabilifying reflective equilibrium. *Synthese*, 203(45).

Dempster, A. P. (1968). A generalization of bayesian inference. *Journal of the Royal Statistical Society*, 30:205–247.

DePaul, M. (2013). Reflective equilibrium. In LaFollette, H., editor, *International Encyclopedia of Ethics*, pages 4466–4475. Wiley, Hoboken, USA.

DePaul, M. R. (1993). *Balance and Refinement: Beyond Coherence Methods of Moral Inquiry*. Routledge, London, UK.

Diamond, I. R., Grant, R. C., Feldman, B. M., Pencharz, P. B., Ling, S. C., Moore, A. M., and Wales, P. W. (2014). Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. *Journal of Clinical Epidemiology*, 67:401–409.

Doorn, N. (2010). Applying rawlsian approaches to resolve ethical issues: Inventory and setting of a research agenda. *Journal of Business Ethics*, 91(1):127–143.

Dorst, K. (2023). Rational polarisation. *Philosophical Review*, 132(3):355–458.

EIU (2023). Democracy index 2023. https://www.eiu.com/n/campaigns/democracy-index-2023/.

Elgin, C. Z. (1996). *Considered Judgment*. Princeton University Press, Princeton, USA.

Elgin, C. Z. (2005). Non-foundationalist epistemology: Holism, coherence, and tenability. In Steup, M. and Sosa, E., editors, *Contemporary Debates in Epistemology*, pages 156–167. Blackwell, Boston, USA.

Elgin, C. Z. (2017). *True Enough*. MIT Press, Cambridge, USA.

Engel, M. (1992). Personal and doxastic justification in epistemology. *Philosophical Studies*, 67(2):133–150.

Enoch, D. (2009). How is moral disagreement a problem for realism? *Journal of Ethics*, 13:15–50.

Espinoza, N. and Peterson, M. (2012). Risk and mid-level moral principles. *Bioethics*, 26(1):8–14.

Estlund, D. and Landemore, H. (2018). The epistemic value of democratic deliberation. In Bächtinger, A., Dryzek, J. S., Mansbridge, J., and Warren,

M., editors, *The Oxford Handbook of Deliberative Democracy*, pages 113–131. Oxford University Press.

Ewing, A. C. (1934). *Idealism: A Critical Survey*. Methuen, London.

Finlayson, J. G. (2019). *The Habermas-Rawls Debate*. Columbia University Press, New York City, USA.

Flick, S. (2022). Ein realistischeres Modell des Überlegungsgleichgewichts. Bachelor's thesis, University of Bern.

Flores, C. and Woodard, E. (2023). Epistemic norms on evidence-gathering. *Philosophical Studies*, 180:2547–2571.

Freeman, S. R. (2007). *Rawls*. Routledge, London, UK.

Freivogel, A. (2021). Modelling reflective equilibrium with belief revision theory. In Blicha, M. and Sedlár, I., editors, *The Logica Yearbook 2020*, pages 65–80. College Publications, London, UK.

Freivogel, A. (2023a). Does reflective equilibrium help us converge? *Synthese*, 202(171).

Freivogel, A. (2023b). *Virtuously Circular: Theoretical Virtues in Reflective Equilibrium*. PhD thesis, University of Bern, Bern, Switzerland.

Freivogel, A. and Cacean, S. (2023). Technical report: Assessing a formal model of reflective equilibrium. Technical report, University of Bern and Karlsruhe Institute of Technology.

Friedman, J. (2020). The epistemic and the zetetic. *The Philosophical Review*, 129(4):501–536.

Gaus, G. (2011). A tale of two sets: Public reason in equilibrium. *Public Affairs Quarterly*, 25(4):305–25.

Genin, K. and Huber, F. (2022). Formal Representations of Belief. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition.

Geuss, R. (2008). *Philosophy and Real Politics*. Princeton University Press, Princeton, USA.

Gigerenzer, G. and Sturm, T. (2012). How (far) can rationality be naturalized? *Synthese*, 187:243–268.

Giugni, M. and Grasso, M., editors (2022). *The Oxford Handbook of Political Participation*. Oxford University Press, Oxford, UK.

Goodman, N. (1955). *Fact, Fiction, & Forecast*. Harvard University Press, Cambridge, USA.

Guardian (2022). Iranian woman dies 'after being beaten by morality police' over hijab law. https://www.theguardian.com/global-development/2022/sep/16/iranian-woman-dies-after-being-beaten-by-morality-police-over-hijab-law [Last accessed: 2023-12-13].

Habermas, J. (1990). *Moral Conciousness and Communicative Action*. MIT Press, Cambridge, USA. Translated by Christian Lenhardt and Shierry Weber Nicholsen.

Habermas, J. (1996). *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT Press, Cambridge, USA.

Hadfield, G. K. and Macedo, S. (2012). Rational reasonableness: Toward a positive theory of public reason. *Law and Ethics of Human Rights*, 6(1):7–46.

Haidt, J., Rosenberg, E., and Hom, H. (2003). Differentiating diversities: Moral diversity is not like other kinds. *Journal of Applied Social Psychology*, 33(1):1–36.

Hájek, A. (2023). Interpretations of Probability. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition.

Hall, B. L., Clover, D. E., Crowther, J., and Scandrett, E., editors (2012). *Learning and Education for a Better World: The Role of Social Movements*. Sense Publishing, Rotterdam, NL.

Hampton, J. (1989). Should political philosophy be done without metaphysics? *Ethics*, 99(4):791–814.

Hegselmann, R. and Krause, U. (2002). Opinion dynamics and bounded confidence: Models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3).

Holey, E. A., Feeley, J. L., Dixon, J., and Whittaker, V. J. (2007). An exploration of the use of simple statistics to measure consensus and stability in Delphi studies. *BMC Medical Research Methodology*, 7(52):1–10.

Hoyningen-Huene, P. (1987). Context of discovery and context of justification. *Studies in History and Philosophy of Science*, 18(4):501–515.

HRW (2023). Human rights report 2023. `https://www.hrw.org/sites/default/files/media_2024/01/World%20Report%202024%20LOWRES%20WEBSPREADS_0.pdf`.

Huber, F. (2013). Belief revision i: The agm theory. *Philosophy Compass*, 8(7):604–612.

Ipsos (2021a). A survey of the American general population (ages 18+). https://de.scribd.com/document/490283649/ABC-News-Ipsos-Poll-Jan-10 [Last accessed: 2023-12-13].

Ipsos (2021b). A survey of the American general population (ages 18+). https://de.scribd.com/document/550570372/ABC-News-Ipsos-Poll-December-27-29-2021 [Last accessed: 2023-12-13].

Iran Human Rights (2023). One year protest report: At least 551 killed and 22 suspicious deaths. https://iranhr.net/en/articles/6200/ [Last accessed: 2023-12-13].

Keefe, R. (2000). *Theories of Vagueness*. Cambridge University Press, Cambridge, UK.

Kelly, T. (2010). Peer disagreement and higher-order evidence. In Feldman, R. and Warfield, T. A., editors, *Disagreement*, pages 111–174. Oxford University Press, Oxford, UK.

Klosko, G. (2015). Rawls, Weithman, and the stability of liberal democracy. *Res Publica*, 21(3):235–249.

Knight, C. (2023). Reflective Equilibrium. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition.

Koehler, D. (2017). *Understanding Deradicalization: Methods, Tools and Programs for Countering Violent Extremism*. Routledge, Abingdon, UK.

Kogelmann, B. (2019). Public reason's chaos theorem. *Episteme*, 16(2):200–219.

Kogelmann, B. and Stich, S. G. W. (2016). When public reason fails us: Convergence discourse as blood oath. *The American Political Science Review*, 110(4):717–730.

Korcz, K. A. (2021). The epistemic basing relation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Kugelberg, H. D. (2021). Public justification versus public deliberation: The case for reconciliation. *Canadian Journal of Philosophy*, 51(6):468–473.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

Küng, H. (1990). *Projekt Weltethos*. Piper, Munich, Germany.

Lambek, S. (2024). The constitutive power of public debate. *Canadian Journal of Political Science*, 57(1):156–173.

Lewis, D. K. (1983). *Philosophical Papers*, volume 1. Oxford University Press, Oxford, UK.

Lin, H. (2024). Bayesian Epistemology. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2024 edition.

Macleod, C. and Tappolet, C., editors (2019). *Philosophical Perspectives on Moral and Civic Education*. Routledge, Abingdon, UK.

Mikhail, J. (2011). *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Jugdment*. Cambridge University Press, Cambridge, UK.

Mills, C. W. (1997). *The Racial Contract*. Cornell University Press, Ithaca, USA.

Mills, C. W. (2005). "Ideal Theory" as ideology. *Hypatia*, 20(3):165–184.

Moehler, M. (2018). *Minimal Morality: A Multilevel Social Contract Theory*. Oxford University Press, Oxford, UK.

Monmouth Poll Reports (2021). One-third remain convinced of 2020 election fraud. `https://www.monmouth.edu/polling-institute/reports/monmouthpoll_us_062121/` [Last accessed: 2023-12-13].

Morton, J. M. (2017). Reasoning under scarcity. *Australasian Journal of Philosophy*, 95(3):543–559.

Müller, M. A. and Campell, J. B. (2023). Arguing in direct democracy: An argument scheme for proposing reasons in debates surrounding public votes. *Topoi*, 42(2):593–607.

Neufeld, B. (2011). Review of 'Why Political Liberalism?'. *Notre Dame Philosophical Reviews*.

NPR (2022). A timeline of how the Jan. 6 attack unfolded — including who said what and when. https://www.npr.org/2022/01/05/1069977469/a-timeline-of-how-the-jan-6-attack-unfolded-including-who-said-what-and-when [Last accessed: 2023-12-13].

NY Times (2020). Trump's attempts to overturn the election are unparalleled in U.S. history. https://edition.cnn.com/2021/11/05/politics/january-6-timeline-trump-coup/index.html [Last accessed: 2023-12-13].

Olsson, E. (2023). Coherentist theories of epistemic justification. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Osborne, M. and Atari, M. (2024). Moral diversity fosters cultural looseness and unpunished norm violations. https://doi.org/10.31234/osf.io/68zwd [Last accessed: 2024-06-19].

Parekh, S. (2020). *No Refuge: Ethics and the Global Refugee Crisis*. Oxford University Press, Oxford, UK.

Parsa, M. (2016). *Democracy in Iran: Why it failed and how it might succeed*. Harvard University Press, Cambridge, USA.

Pateman, C. (1988). *The Sexual Contract*. Polity Press, Cambridge, UK.

Peterson, A., Stahl, G., and Soong, H., editors (2020). *The Palgrave Handbook of Citizenship and Education*. Palgrave Macmillan Cham, London, UK.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford University Press, Oxford, UK.

Politifact (2020). List does not show over 14,000 dead people cast ballots in Michigan's Wayne County. https://www.politifact.com/factchecks/2020/nov/07/tweets/list-does-not-show-over-14000-dead-people-cast-bal/ [Last accessed: 2020-12-13].

Popper, K. (2012). *The Open Society and Its Enemies*. Routledge, Abingdon, UK.

Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review*, 60(1):20–43.

Quong, J. (2011). *Liberalism without Perfection*. Oxford University Press, Oxford, UK.

Quong, J. (2022). Public Reason. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2022 edition.

Rabasa, A., Pettyjohn, S. L., Ghez, J. J., and Boucek, C. (2010). *Deradicalizing Islamist Extremists*. RAND, Santa Monica, USA.

Rasmussen Reports (2022). Election integrity will be important issue in November, voters say. https://www.rasmussenreports.com/public_content/politics/general_politics/march_2022/election_integrity_will_be_important_issue_in_november_voters_say?utm_campaign=RR03082022DN&utm_source=criticalimpact&utm_medium=email [Last accessed: 2023-12-13].

Rawls, J. (1974). The independence of moral theory. *Proceedings and Addresses of the American Philosophical Association*, 48:5–22.

Rawls, J. (1985). Justice as fairness: Political not metaphysical. *Philosophy & Public Affairs*, 14(3):223–251.

Rawls, J. (1987). The idea of an overlapping consensus. *Oxford Journal of Legal Studies*, 7(1):1–25.

Rawls, J. (1997). The idea of public reason revisited. *The University of Chicago Law Review*, 64(3):765–807.

Rawls, J. (1999). *A Theory of Justice*. Belknap Press, Cambridge, USA, revised edition.

Rawls, J. (2001). *Justice as Fairness: A Restatement*. Belknap Press, Cambridge, UK.

Rawls, J. (2005). *Political Liberalism*. Columbia University Press, New York City, USA, expanded edition.

Raz, J. (1990). Facing diversity: The case of epistemic abstinence. *Philosophy & Public Affairs*, 19(1):3–46.

Rechnitzer, T. (2022). *Applying Reflective Equilibrium: Towards the Justification of a Precautionary Principle*. Springer, Cham, Switzerland.

Reichenbach, H. (1938). *Experience and Prediction*. University of Chicago Press, Chicago, USA.

Reuters (2023). What has changed in Iran one year since mahsa amini protests erupted? https://www.reuters.com/world/middle-east/what-has-changed-iran-one-year-since-mahsa-amini-protests-erupted-2023-09-11/.

Rogers, T. (2020). Virtue ethics and political authority. *Journal of Social Philosophy*, 51:303–321.

Scanlon, T. M. (2003). Rawls on justification. In Freeman, S. R., editor, *The Cambridge Companion to Rawls*, pages 139–167. Cambridge University Press, Cambridge, UK.

Scanlon, T. M. (2014). *Being Realistic about Reasons*. Oxford University Press, Oxford, UK.

Schmidt, M. W. (2022). *Das Überlegungsgleichgewicht als Lebensform: Versuch zu einem vertieften Verständnis der durch John Rawls bekannt gewordenenen Rechtfertigungsmethode*. Brill Mentis, Paderborn, Germany.

Shannon, C. E. (1998). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

Shieber, J. (2015). *Testimony: A Philosophical Introduction*. Routledge, London, UK.

Shogenji, T. (1999). Is coherence truth-conducive? *Analysis*, 59:338–345.

Silber, L. and Little, A. (1996). *The Death of Yugoslavia*. Penguin, London, UK, revised edition.

Simon, H. A. (1957). *Models of Man*. Wiley, New York City, USA.

Singer, D. J., Bramson, A., Grim, P., Holman, B., Jung, J., Kovaka, K., Ranginani, A., and Berger, W. J. (2019). Rational social and political polarization. *Philosophical Studies*, 176(9):2243–2269.

Spohn, W. (2012). *The Laws of Belief: Ranking Theory and its Philosophical Applications*. Oxford University Press, Oxford, UK.

Stahl, T. (2022). What (if anything) is ideological about ideal theory? *European Journal of Political Theory*, 23(2):135–158.

Stanley, R. P. (1997). *Enumerative Combinatorics*, volume 1. Cambridge University Press, Cambridge, UK.

Stemmer, P. (2000). *Handeln zugunsten anderer. Eine moralphilosophische Untersuchung.* de Gruyter, Berlin, Germany.

Strasser, C. and Antonelli, G. A. (2024). Non-monotonic Logic. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2024 edition.

Swanton, C. (1992). *Freedom: A Coherence Theory*. Hackett Publishing, Indianapolis, USA.

Taylor, C. (1999). Conditions of an unforced consensus on human rights. In Bauer, J. R. and Bell, D. A., editors, *The East Asian Challenge for Human Rights*, pages 124–145. Cambridge University Press.

Tersman, F. (1993). *Reflective Equilibrium: An Essay in Moral Epistemology*. Almqvist & Wiksell, Stockholm, Sweden.

Tersman, F. (2006). *Moral Disagreement*. Cambridge University Press, Cambridge, UK.

Thagard, P. (2000). *Coherence in Thought and Action*. MIT Press, Cambridge, USA.

Thorstad, D. (2022). There are no epistemic norms of inquiry. *Synthese*, 200(410):1–24.

Thorstad, D. (2023). Why bounded rationality (in epistemology)? *Philosophy and Phenomenological Research*, 108(2):369–413.

Thrasher, J. and Vallier, K. (2013). The fragility of consensus: Public reason, diversity and stability. *European Journal of Philosophy*, 23(4):933–954.

Turri, J. (2010). On the relationship between propositional and doxastic justification. *Philosophy and Phenomenological Research*, 80(2):312–326.

Valentini, L. (2012). Ideal vs. non-ideal theory: A conceptual map. *Philosophy Compass*, 7(9):654–664.

Vallier, K. (2015). Public justification versus public deliberation: The case for divorce. *Canadian Journal of Philosophy*, 45(2):139–158.

Vallier, K. (2022). Public Justification. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition.

van der Burg, W. and van Willigenburg, T., editors (1998). *Reflective Equilibrium: Essays in Honor of Robert Heeger*. Kluwer, Alphen, Netherlands.

Walzer, M. (1997). *On Toleration*. Yale University Press, New Haven, USA.

Washington Post (2021). What Trump said before his supporters stormed the Capitol, annotated. https://www.washingtonpost.com/politics/interactive/2021/annotated-trump-speech-jan-6-capitol/ [Last accessed: 2023-12-13].

Weithman, P. (2010). *Why Political Liberalism?* Oxford University Press, Oxford, UK.

Weithman, P. (2015). Inclusivism, stability, and assurance. In Bailey, T. and Gentile, V., editors, *Rawls and Religion*, pages 75–96. Columbia University Press, New York City, USA.

Weithman, P. (2023). Stability and equilibrium in political liberalism. *Philosophical Studies*, 181:23–41.

Wenar, L. (2021). John Rawls. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition.

White, R. (2005). Epistemic permissiveness. *Philosophical Perspectives*, 19:445–459.

Williams, B. (2005). *In the Beginning Was the Deed: Realism and Moralism in Political Argument*. Princeton University Press, Princeton, USA.

Wong, B. and Li, M.-K. (2023). Talk may be cheap, but deeds seldom cheat: On political liberalism and the assurance problem. *American Journal of Political Science*, n/a(n/a).

Appendix

A The achievement function

The achievement function is a measure for the degree to which an epistemic state is in reflective equilibrium (see section 3.2). It is the weighted sum of faithfulness, account and systematicity. Let's start with account: The basic idea is that a theory accounts for a sentence iff the sentence is in the theory's content. Regarding sets of sentences, like an agent's commitments, account is a matter of degree. Let (S, A) be a dialectical structure. A theory T accounts for some commitments C to the extent that its content $\overline{T}$ contains all of them and only them. Formally, we first measure the (weighted) Hamming distance between C and $\overline{T}$, and normalise it by the number of unnegated sentences $N := |S|/2$. Then, we plug this distance into a monotonically decreasing function $G : \mathbb{R} \to \mathbb{R}$ to get the "closeness" of C to $\overline{T}$. This closeness is the degree to which T accounts for C:

$$A(C, T) := G\Big(\frac{D_{0,0.3,1,1}\left(C, \overline{T}\right)}{N}\Big),$$

where D is a weighted Hamming Distance between arbitrary positions P, Q,

$$D_{d_0,d_1,d_2,d_3}(P, Q) := \sum_{\{s,\neg s\}\subset S} d_{d_0,d_1,d_2,d_3}(P, Q, \{s, \neg s\})$$

with the penalty function d:

$$d_{d_0,d_1,d_2,d_3}(P,Q,\{s,\neg s\}) = \begin{cases} d_3 & \text{if} \quad \{s,\neg s\} \subset (P \cup Q) \\ d_2 & \text{if} \quad \{s,\neg s\} \cap P \neq \emptyset \wedge \{s,\neg s\} \cap Q = \emptyset \\ d_1 & \text{if} \quad \{s,\neg s\} \cap P = \emptyset \wedge \{s,\neg s\} \cap Q \neq \emptyset \\ d_0 & \text{otherwise.} \end{cases} .$$

Thus far, the model is tested for two monotonically decreasing functions G, a quadratic and a linear one:

$$G_{quadratic}(x) := 1 - x^2,$$
$$G_{linear}(x) := 1 - x.$$

Faithfulness is supposed to be a tie to the initial commitments. Its functional representation measures how close the commitments are to the initial commitments. The definition is almost completely analogous to the previous one. The only difference is that extending the initial commitments is not penalised:

$$F(C|C_0) := G\Big(\frac{D_{0,0,1,1}(C_0, C)}{N}\Big).$$

Lastly, a theory's systematicity is a combination of its simplicity and scope. The number of the theory's principles (minus a parameter ξ) is divided by the number of sentences in its content. The result is plugged into the monotonically decreasing function G:

$$S(T) := G\Big(\frac{|T| - \xi}{|\overline{T}|}\Big),$$

with $\xi := 0.99$ and $S(\emptyset) := 0$. As a result, $S(T)$ increases with less principles and more content, as desired. If we did not subtract ξ from the number of principles, then a systematicity close to 1 would be impossible, even for a single principle with maximal content $|\overline{T}| = |S|/2$. Why not choose $\xi = 1$ such that systematicity can be 1, like the other desiderata? If $\xi = 1$, then all theories with $|T| = 1$ have the same systematicity of 1, i.e. their content does not make a difference anymore. With $\xi = 0.99$, we can differentiate between these single-principle-theories whilst still getting systematicity values that

are very close to 1. (In the BBB model as published by Beisbart et al. (2021), $\xi = 1$. Thus, setting $\xi = 0.99$ is a (very) slight deviation from their original model.)

Now we trade off these desiderata to obtain an epistemic state's achievement:

$$Z(C, T|C_0) := \alpha_F \cdot F(C|C_0) + \alpha_A \cdot A(T, C) + \alpha_S \cdot S(T),$$

with non-negative weights $\alpha_F, \alpha_A, \alpha_S$ adding up to 1. As a standard configuration, $\alpha_F = 0.1,\ \alpha_A = 0.35,\ \alpha_S = 0.55$ has proven to provide plausible results. For more on this, see Beisbart, Betz and Brun's 2021 paper.

B Entropy and Kullback-Leibler divergence

Definition 11 (KL divergence)**.** Let q and r be probability distributions over Opt_{CD} such that $r(O) = 0$ only if $q(O) = 0$ for all $O \in Opt_{CD}$. (A probability distribution over a finite set of mutually exclusive and jointly exhaustive outcomes is a function from that set to non-negative real values summing to 1.) Let $b \in \mathbb{R}$ with $0 < b \neq 1$. The Kullback-Leibler divergence of q from r with base b is defined as

$$D^b_{KL}(q||r) := \sum_{O \in Opt_{CD}} q(O)\, log_b \left(\frac{q(O)}{r(O)}\right),$$

where any term of this sum is 0 if $q(O) = 0$. (I adapted the original definition by Kullback and Leibler (1951) to the present scenario.)

Proposition 1. Let $((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m}))$ (with $1 < m =: n_{FP}$) be a subtuple of $((C_1, T_1), \ldots, (C_n, T_n)) \in \mathfrak{E}$. For $O \in Opt_{CD}$ let $p(O) := |\{C_{s_i} : C_{s_i} \text{ realises } O\}|/n_{FP}$ which is a probability distribution over Opt_{CD}. Let $Opt^{p>0}_{CD} := \{O \in Opt_{CD} : p(O) > 0\}$. Let u be one of the most uniform probability distributions over Opt_{CD} that are achievable given the length n_{FP} of the subtuple. That is, u is a probability distribution such that $u(O) = 1/|Opt_{CD}|$ for all $O \in Opt_{CD}$, unless $n_{FP} < |Opt_{CD}|$, in which case $u(O) = 1/n_{FP}$ for exactly n_{FP} options $O \in Opt_{CD}$ such that $u(O) = 0$ only if $p(O) = 0$ for all $O \in Opt_{CD}$. (For a brief explanation of the construction of u, see the remark below the proof.)

Then there are $b, c \in \mathbb{R}$ such that

$$c \cdot Entropy((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m})) = 1 - D^b_{KL}(p||u).$$

Proof. Let $b = max := min(|Opt_{CD}|, n_{FP})$ and $c = 1/100$. Then the following are equivalence transformations (each step is explained below):

$$\begin{aligned}
D^{max}_{KL}(p||u) &= \sum_{O \in Opt_{CD}} p(O)\ log_{max}\left(\frac{p(O)}{u(O)}\right) && (6.1)\\
&= \sum_{O \in Opt^{p>0}_{CD}} p(O)\ log_{max}\left(\frac{p(O)}{u(O)}\right) && (6.2)\\
&= \sum_{O \in Opt^{p>0}_{CD}} p(O)\ log_{max}\left(\frac{p(O)}{1/max}\right) && (6.3)\\
&= \sum_{O \in Opt^{p>0}_{CD}} p(O)\ \ (log_{max} p(O) + log_{max} max) && (6.4)\\
&= \sum_{O \in Opt^{p>0}_{CD}} p(O)\ \ (log_{max} p(O) + 1) && (6.5)\\
&= \sum_{O \in Opt^{p>0}_{CD}} p(O)\ log_{max} p(O) + p(O) && (6.6)\\
&= \sum_{O \in Opt^{p>0}_{CD}} p(O) + \sum_{O \in Opt^{p>0}_{CD}} p(O)\ log_{max} p(O) && (6.7)\\
&= 1 + \sum_{O \in Opt^{p>0}_{CD}} p(O)\ log_{max} p(O) && (6.8)\\
&= 1 + \sum_{O \in Opt_{CD}} p(O)\ log_{max} p(O) && (6.9)\\
&= 1 - \frac{1}{100} \cdot 100 \cdot (-1) \sum_{O \in Opt_{CD}} p(O)\ log_{max} p(O) && (6.10)\\
&= 1 - \frac{1}{100} \cdot Entropy((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m})) && (6.11)
\end{aligned}$$

which is equivalent to

$$c \cdot Entropy((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m})) = 1 - D^b_{KL}(p||u)$$

which was to be shown. Explanation of these steps: (1) by definition of KL

divergence; (2) because any term of the sum is 0 if $p(O) = 0$ (by definition of KL divergence); (3) because $u(O) = 0$ only if $p(O) = 0$ for all $O \in Opt_{CD}$, and other than that u can return only $1/|Opt_{CD}|$ or $1/n_{FP}$, depending on whether $n_{FP} < |Opt_{CD}|$ or not, respectively; (4) by known logarithmic identities; (5) by known logarithmic identities; (6) by expansion of product; (7) by commutativity of addition; (8) because p is a probability distribution, thus, its (non-zero) values sum to 1; (9) because any term of the sum is 0 if $p(O) = 0$; (10) trivial; (11) by definition of *Entropy* (see section 3.4.2).

Remark: If the subtuple is long enough, u is simply the uniform distribution over the options. However, if there are less fixpoints in the subtuple than options ($n_{FP} < |Opt_{CD}|$), then the uniform distribution is not achievable since there will always be options $O \in Opt_{CD}$ with $p(O) = 0$. But intuitively, there is a set of 'most uniform distributions': Those according to which every fixpoint realises a different option. Thus, the KL divergence is measured from one of these. It doesn't really matter which one (any will give the same result) as long as we choose one such that $u(O) = 0$ only if $p(O) = 0$ for all $O \in Opt_{CD}$ (otherwise KL divergence is not well-defined).

C Collection of most important explications

I here restate the most important definitions and explications from chapter 3.

Consensus

Let $A = \{a_1, ..., a_n\}$ be a set of agents living together in a society and sharing the dialectical structure (S, A). Let $\mathcal{E} = \mathcal{C} \times \mathcal{T}$ be the set of possible epistemic states with the set of all minimally consistent positions $\mathcal{C} \subset \wp(S)$ (the possible commitments) and the set of all dialectically consistent positions $\mathcal{T} \subset \wp(S)$ (the possible theories). Let $\mathcal{J}^{a_i} \subset \mathcal{E}$ be the set of epistemic states justified for agent a_i and $\mathfrak{E} := \mathcal{J}^{a_1} \times \cdots \times \mathcal{J}^{a_n}$ the space of justified belief systems for their society. Let $PC \in S$ be a political conception of justice.

Definition 6 (Acceptance rate)**.** The *acceptance rate of PC* in a tuple

$((C_1, T_1), \ldots, (C_n, T_n)) \in \mathfrak{E}$ is given by:

$$AccRate((C_1, T_1), \ldots, (C_n, T_n)) = \frac{100}{n}|\{C_i : PC \in C_i\}|$$

Pluralism

Let $A = \{a_1, ..., a_n\}$ be a set of agents living together in a society and sharing the dialectical structure (S, A). Let $\mathcal{E} = \mathcal{C} \times \mathcal{T}$ be the set of possible epistemic states with the set of all minimally consistent positions $\mathcal{C} \subset \wp(S)$ (the possible commitments) and the set of all dialectically consistent positions $\mathcal{T} \subset \wp(S)$ (the possible theories). Let $\mathcal{J}^{a_i} \subset \mathcal{E}$ be the set of epistemic states justified for agent a_i and $\mathfrak{E} := \mathcal{J}^{a_1} \times \cdots \times \mathcal{J}^{a_n}$ the space of justified belief systems for their society. Let $PC \in S$ be a political conception of justice.

Definition 7 (Subtuple and *PC*-subtuple)**.** Let $\mathfrak{T} = ((C_1, T_1), \ldots, (C_n, T_n)) \in \mathfrak{E}$. Let $I \subset \{1, \ldots, n\}$ with cardinality $m := |I| = n_{FP}$ and elements $s_1 < ... < s_m$. Then $((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m}))$ is a *subtuple* of $\mathfrak{T}$. In particular, let $I_{PC} := \{i \in \{1, \ldots, n\} : PC \in C_i\}$ with cardinality $m := |I_{PC}|$ and elements $s_1 < ... < s_m$. Then the *PC-subtuple* of $\mathfrak{T}$ is defined as

$$\mathfrak{T}^{PC} := ((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m})).$$

For a given subtuple, the maximum number of realisable CD-options is given by $max := \min(\{\ n_{CD} + 1,\ n_{FP}\ \})$ with the number comprehensive doctrines n_{CD} in the structure and the length of the subtuple n_{FP}.

Definition 8 (Option count)**.** The *option count* of a subtuple $((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m}))$ (with $m = n_{FP} \leq n$) of some tuple $((C_1, T_1), \ldots, (C_n, T_n)) \in \mathfrak{E}$ is given by

$$OptCount((C_{s_1}, T_{s_1}), \ldots, (C_{s_n}, T_{s_n})) := 100 \cdot \frac{|\{O \in Opt_{CD} : \exists C_{s_i} \text{ s.t. } C_{s_i} \text{ realises } O\}| - 1}{max - 1}.$$

Definition 9 (Strength of the weak)**.** Let $((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m}))$ (with $m = n_{FP} \leq n$) be a subtuple of $((C_1, T_1), \ldots, (C_n, T_n)) \in \mathfrak{E}$. Let $O_{strong} \in Opt_{CD}$ be the strongest CD-option (or one of them if there are several), i.e. the CD-option that is realised in the commitments of at least as many epistemic

states in the subtuple as any other CD-option. The *strength of the weak* of $((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m}))$ is given by

$$SoW((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m})) := 100 \cdot \frac{|\{C_{s_i} : C_{s_i} \text{ does not realise } O_{strong}\}|/n_{FP}}{(max - 1)/max}.$$

Definition 10 (Entropy)**.** Let $((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m}))$ (with $m = n_{FP} \leq n$) be a subtuple of $((C_1, T_1), \ldots, (C_n, T_n)) \in \mathfrak{E}$. For $O \in Opt_{CD}$ let $p(O) := |\{C_{s_i} : C_{s_i} \text{ realises } O\}|/n_{FP}$. The *entropy* of $((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m}))$ is given by

$$Entropy((C_{s_1}, T_{s_1}), \ldots, (C_{s_m}, T_{s_m})) := 100 \cdot (-1) \sum_{O \in Opt} p(O) \log_{max} p(O).$$

Different kinds of overlapping consensus

Let $A = \{a_1, ..., a_n\}$ be a set of agents living together in a society with a shared dialectical structure (S, A). Let $\mathcal{E} = \mathcal{C} \times \mathcal{T}$ be the set of possible epistemic states with the set of all minimally consistent positions $\mathcal{C} \subset \wp(S)$ (the possible commitments) and the set of all dialectically consistent positions $\mathcal{T} \subset \wp(S)$ (the possible theories). Let $F((S, A), C_0, Alg) \subset \mathcal{E}$ be the set of all possible fixed points (in particular, considering all random choices) of algorithm Alg applied to initial commitments C_0 on dialectical structure (S, A). Let $C_0^{a_i} \in \mathcal{C}$ be a_i's initial commitments. Let $t_{AccRate}$ be the lowest plausible threshold for categorical consensus when using the acceptance rate as a measure for gradual consensus, likewise $t_{OptCount}$, t_{SoW} and $t_{Entropy}$ for the corresponding pluralism measures. Let *Pluralism* be a variable that can take three values: *OptCount*, *SoW* and *Entropy*. Let *LocalMRE* be a variable that can take two values: *LocalQuadraticMRE* and *LocalLinearMRE*. Let $PC \in S$ be a political conception of justice. Then:

Explication 3 (Potential global overlapping consensus)**.** There is a *potential global overlapping consensus on PC*

- *in the weak sense* iff there is at least one tuple

$\mathfrak{T} \in F((S,A), C_0^{a_1}, LocalMRE) \times \cdots \times F((S,A), C_0^{a_n}, LocalMRE)$ with

$$AccRate(\mathfrak{T}) \geq t_{AccRate}, \text{ and}$$
$$Pluralism(\mathfrak{T}^{PC}) \geq t_{Pluralism}.$$

- *in the strong sense* iff for all tuples $\mathfrak{T} \in F((S,A), C_0^{a_1}, LocalMRE) \times \cdots \times F((S,A), C_0^{a_n}, LocalMRE)$:

$$AccRate(\mathfrak{T}) \geq t_{AccRate}, \text{ and}$$
$$Pluralism(\mathfrak{T}^{PC}) \geq t_{Pluralism}.$$

- *of grade r* iff for a proportion $r \in [0,1]$ of all tuples $\mathfrak{T} \in F((S,A), C_0^{a_1}, LocalMRE) \times \cdots \times F((S,A), C_0^{a_n}, LocalMRE)$, it holds that

$$AccRate(\mathfrak{T}) \geq t_{AccRate}, \text{ and}$$
$$Pluralism(\mathfrak{T}^{PC}) \geq t_{Pluralism}.$$

Explication 4 (Potential local overlapping consensus)**.** There is a *potential local overlapping consensus on PC*

- *in the weak sense* iff there is at least one tuple $\mathfrak{T} \in F((S,A), C_0^{a_1}, LocalMRE) \times \cdots \times F((S,A), C_0^{a_n}, LocalMRE)$ with

$$Pluralism(\mathfrak{T}^{PC}) \geq t_{Pluralism}.$$

- *in the strong sense* iff for all tuples $\mathfrak{T} \in F((S,A), C_0^{a_1}, LocalMRE) \times \cdots \times F((S,A), C_0^{a_n}, LocalMRE)$:

$$Pluralism(\mathfrak{T}^{PC}) \geq t_{Pluralism}.$$

- *of grade r* iff for a proportion $r \in [0,1]$ of all tuples $\mathfrak{T} \in F((S,A), C_0^{a_1}, LocalMRE) \times \cdots \times F((S,A), C_0^{a_n}, LocalMRE)$, it holds that

$$Pluralism(\mathfrak{T}^{PC}) \geq t_{Pluralism}.$$

D Results for different political conceptions

In section 5.1, I said that there is no harm in arbitrarily fixing PC1 as the political conception regarding which we test the research hypotheses. The results will be the same (or very similar) for PC2 and PC3, because the study is completely symmetric regarding the PCs. In this section of the appendix, I show that this holds by presenting the ternary heatmaps of the average acceptance rates for PC1, PC2 and PC3. (I chose the average acceptance rates as an example, the same holds for all other values, e.g. the pluralism values.) As you can see in figures 6.1–6.3, the ternaries look exactly the same. Of course, PC2 does not occur in societies with $n_{PC} = 1$ and PC3 only occurs in societies with $n_{PC} = 3$. But when comparing ternaries where a comparison is possible, the values are mostly identical. Sometimes they differ by 1 or so, but this is to be expected. Identical averages are guaranteed only if we simulate possibility space as a whole.

E Acceptance mechanisms

In section 5.1.2 I distinguished the acceptance mechanisms M1–3 and pointed out that they are exhaustive, i.e. there is no other way that PC1 can end up in the fixpoint commitments. This fact is explained by the following considerations:

- First of all, it is not possible that PC1 is accepted in the fixpoint commitments even though it is not implied by the fixpoint theory. This follows from the fact that PC1 does not occur in the initial commitments to begin with. The only incentive to add it later can be to increase account because the theory implies it. Given that PC1 is implied by the theory (if it's accepted in the fixpoint commitments at all), then there are two cases: Either PC1 is *not* in the theory (but still implied by it) or it *is* in the theory (and trivially implied by it).

- If PC1 is *not* in the theory, it can only be implied by a theory with a supportive CD. This follows from the general fact that no combination of particular statements can imply the general statement PC1, and the other PCs are incompatible with PC1 by design. Thus, a CD must be in

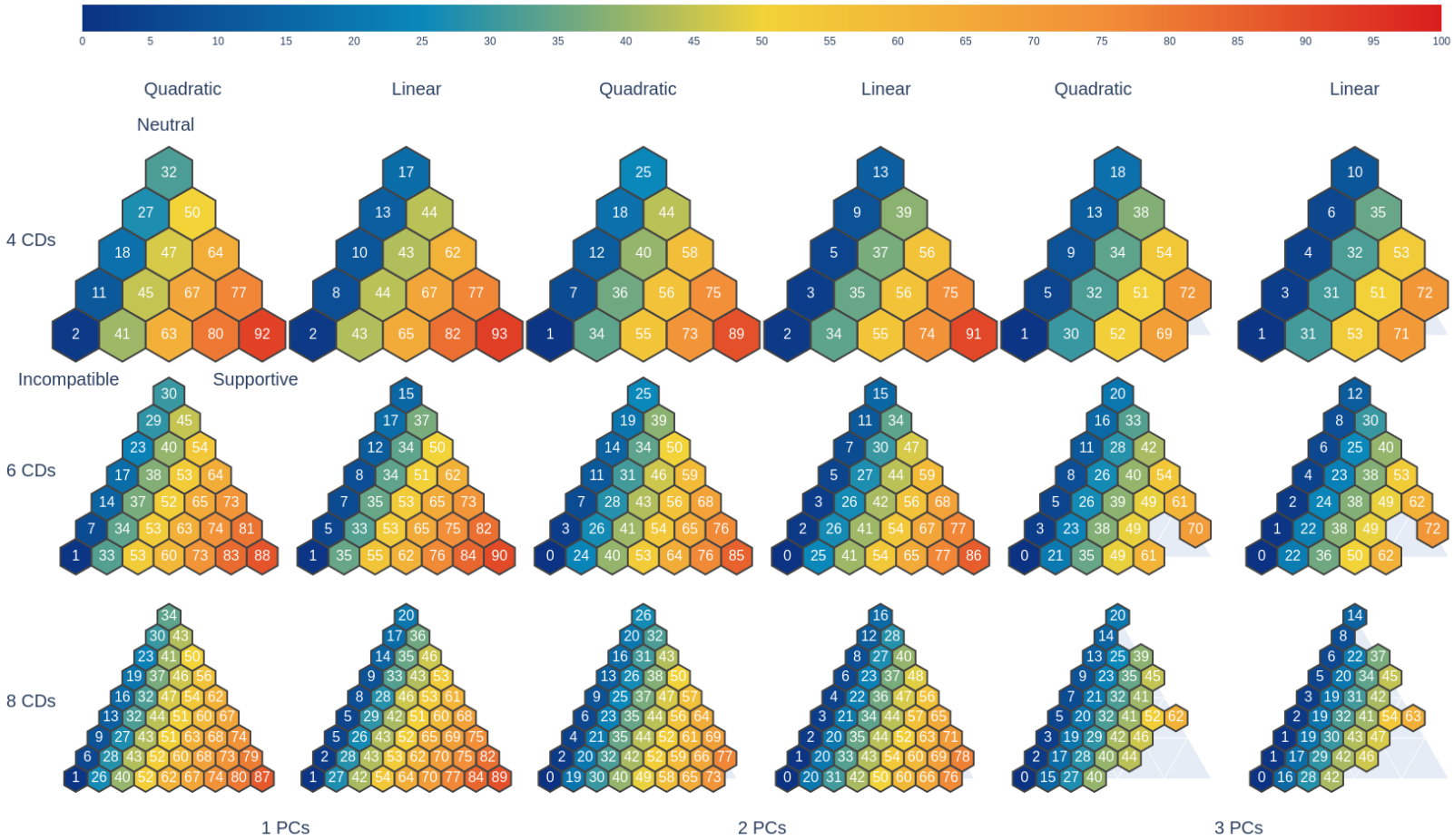

Figure 6.1: These ternary plots display the arithmetic averages of the *acceptance rates* for PC1. The data is split up according to model variant, n_{CD} and n_{PC}.

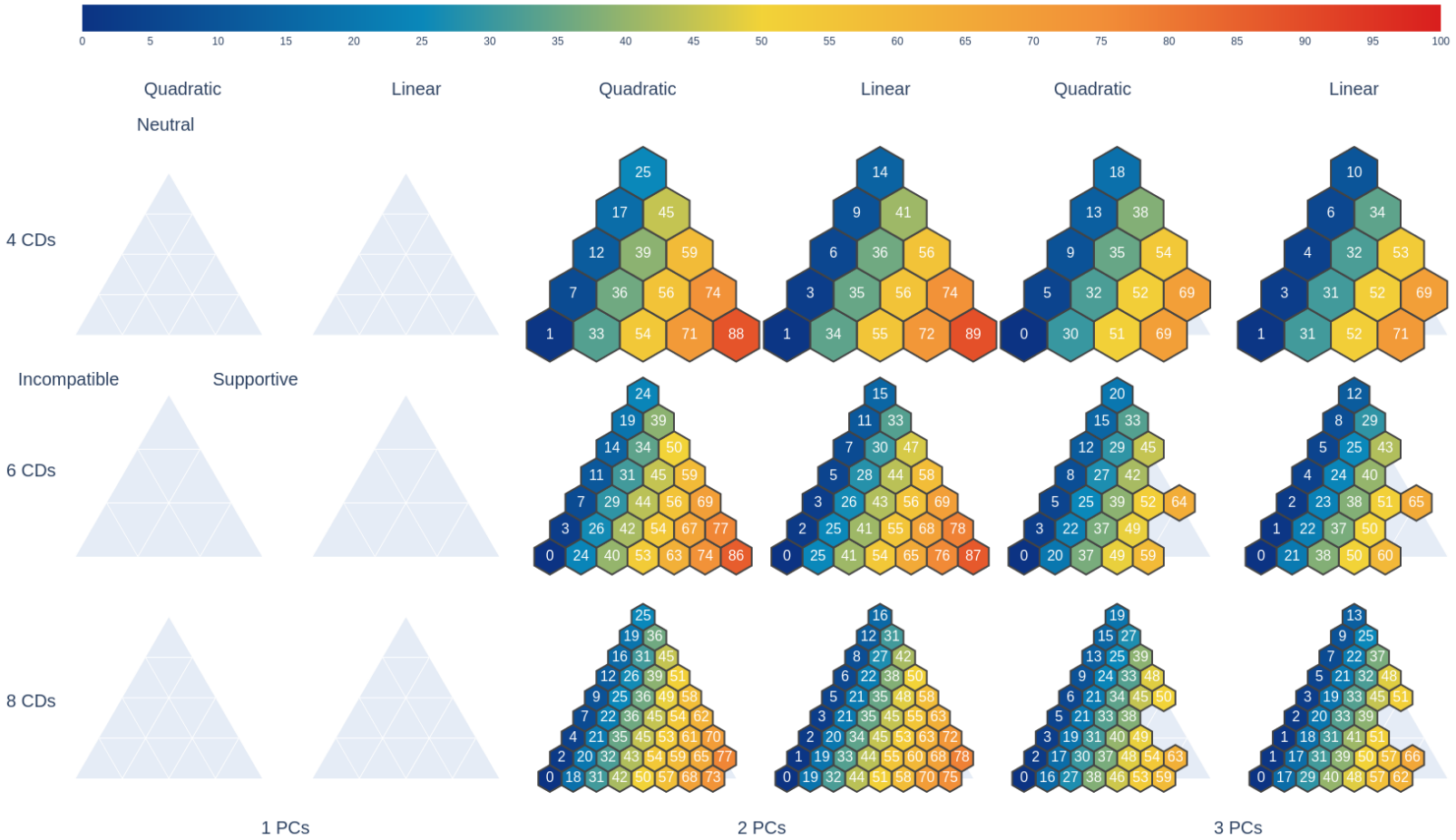

Figure 6.2: These ternary plots display the arithmetic averages of the *acceptance rates* for PC2. The data is split up according to model variant, n_{CD} and n_{PC}.

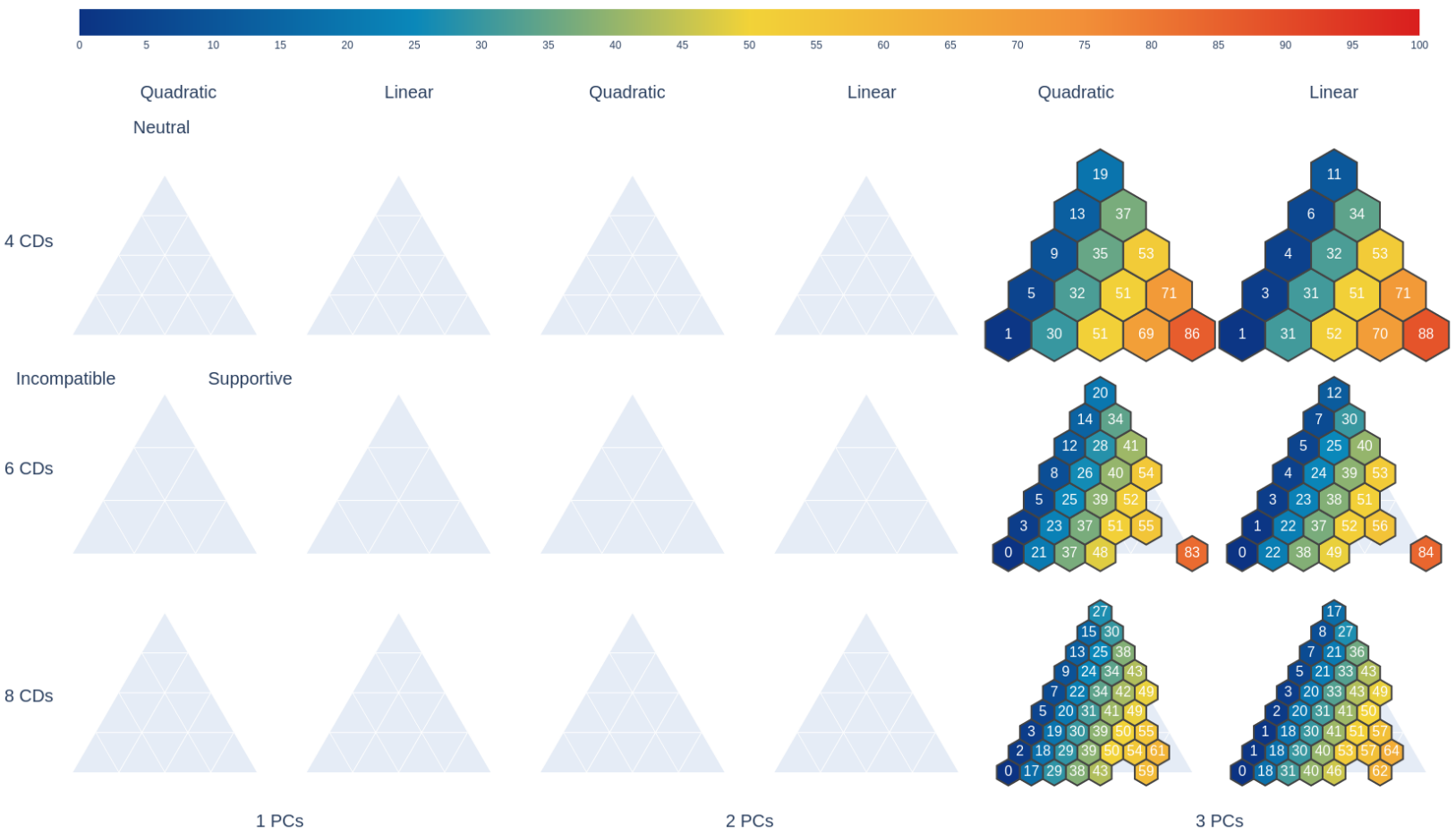

Figure 6.3: These ternary plots display the arithmetic averages of the *acceptance rates* for PC3. The data is split up according to model variant, n_{CD} and n_{PC}.

the theory. But of the CDs (even in combination with other sentences) only one with an s-connection can imply PC1. This is mechanism M1.

- If PC1 *is* in the theory, then either together with a neutral CD or no CD at all. This follows from the fact that PC1 together with an incompatible CD would render the theory inconsistent. PC1 together with a supportive CD would decrease systematicity without improving account: We could improve achievement during theory adjustment by removing PC1 from the theory. This would improve systematicity without a loss in account, because a theory with a supportive CD implies PC1 anyways. Thus, PC1 will not occur in the theory together with a supportive CD. On the bottom line, if PC1 is in the theory, then only together with a neutral CD (constituting M2) or no CD at all (constituting M3).

These considerations explain why M1–3 are exhaustive.

F Global pluralism

Some arguments in section 5.1.3 relied on the assumption that globally speaking, i.e. in societies as a whole, the CD-options are realised more or less with a homogeneous distribution. This assumption is here backed up by showing that the global entropy (being a measure for the homogeneity of the distribution) is uniformly high throughout the ternaries, as you can see in figure 6.4.

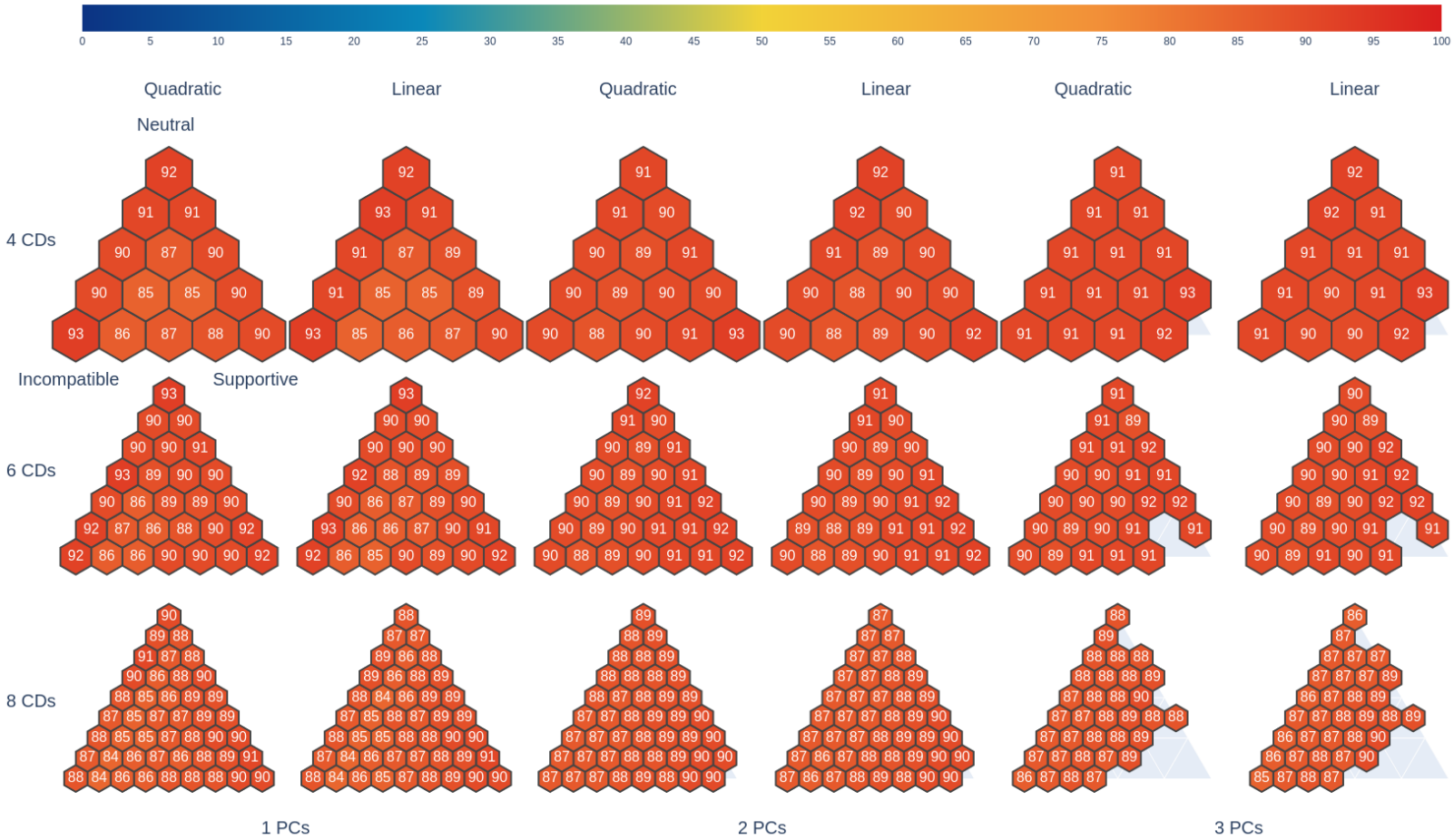

Figure 6.4: These ternary plots display the arithmetic averages of the *global entropy*, i.e. the entropy off all fixpoints, not just the ones that accept PC1. The data is split up according to model variant, n_{CD} and n_{PC}.