Virtuously Circular

Theoretical Virtues in Reflective Equilibrium

Inauguraldissertation
an der Philosophisch-historischen Fakultät der Universität Bern
zur Erlangung der Doktorwürde

vorgelegt von

Andreas Freivogel

Promotionsdatum: 13.10.2023

eingereicht bei

Prof. Dr. Dr. Claus Beisbart
Institut für Philosophie, Universität Bern
und
Prof. Dr. Georg Brun
Institut für Philosophie, Universität Bern

Acknowledgements

The conception and completion of my dissertation would not have been possible without the continuous support of Claus Beisbart and Georg Brun, to whom I want to express my sincere gratitude. They infected me with their enthusiasm for reflective equilibrium and its formal-computational study long before the project started. As supervisors, they provided me with invaluable guidance. Their ingenious suggestions and constructive criticism led to significant improvements in all aspects of my thesis.

I am also extremely grateful to Gregor Betz for sharing his deep insights of the formal framework, which benefited the conception of simulation studies and the interpretation of results. A visit in Karlsruhe marked the inspiring starting point to the development of the computer program.

I would like to extend my sincere thanks to Sebastian Cacean, Richard Lohse, and Alexander Koch. In dozens of meetings as a genuine project team, I had the privilege to present, discuss and sharpen my ideas on many occasions. The help that I received from all of you cannot be overestimated.

Special thanks go to Sebastian Cacean. It is thanks to him that the computer code, to which I contributed, is well structured, documented and tested.

I also wish to thank Tanja Rechnitzer. My work profited from the insightful exchange of ideas, and from her great example as my predecessor at Bern.

I gratefully acknowledge the instructive opportunity to provide technical support for Sebastian Flick and Noah Werder in their projects working with the computer program.

Last but not least, a heartfelt thanks go to my family for their unwavering support and patience. I'm deeply indebted to Michaela, my wife. Thank you for keeping me sane during this time!

This work was funded by the Swiss National Science Foundation as part of the SNSF-DFG project "How far does reflective equilibrium take us? Investigating the power of a philosophical method" (Swiss National Science Foundation grant 182854 and German Research Foundation grant 412679086).

Contents

To Irina and Aurel

Chapter 1

Introduction

1.1 Reflective Equilibrium

The title of this dissertation – "Virtuously Circular" – draws inspiration from Nelson Goodman's classic exhibition of an account of justification:

> Principles of deductive inference are justified by their conformity with accepted deductive practice. Their validity depends upon accordance with the particular deductive inferences we actually make and sanction. If a rule yields inacceptable inferences, we drop it as invalid. Justification of general rules thus derives from judgments rejecting or accepting particular deductive inferences.
>
> This looks flagrantly circular. I have said that deductive inferences are justified by their conformity to valid general rules, and that general rules are justified by their conformity to valid inferences. But this circle is a virtuous one. The point is that rules and particular inferences alike are justified by being brought into agreement with each other. *A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend*. The process of justification is the delicate one of making mutual adjustments between rules and accepted inferences; and in the agreement achieved lies the only justification needed for either. (Goodman, 1983(1955), 63–64, emphasis in original)

Notably, John Rawls, who coined the term "reflective equilibrium", and who brought the idea into prominence in *A Theory of Justice*, refers to this passage (1971, 20). Above quote is a good entry point as it highlights the key ideas about reflective equilibrium (henceforth: RE).

First, RE addresses the age-old problem of *justification*, which is central to the philosophical discipline called *epistemology*. Justification, classically understood, is supposed to differentiate between genuine knowledge and mere opinion that happens to be true. Goodman alludes with his "virtuous circle" to the idiomatic "vicious circle", which is part of the problem of justification. The problem arises in the form of the so-called *Münchhausen Trilemma* as described by Albert (1985, 18): If everything stands in need of justification, whatever we put forward to provide justification, needs to be justified as well. This process of providing justification either i) goes back further and further in an *infinite regress*, ii) results in a *logical circle*, where two parts lend support to each other in a reciprocal manner, or iii) it breaks off by *recourse on a dogma*. Typically, neither horn of the trilemma appears to be acceptable at first sight.

The quote from above suggests that Goodman opts for the second horn of the trilemma. Goodman notes that his solution "looks flagrantly circular", but he is quick to respond that the circle is "virtuous". This is a typically coherentist strategy. The idea is that coherence, Goodman calls it "agreement", is not a vicious form of circularity. Roughly, a body of elements is coherent if the elements are consistent with each other and if they are mutually supportive. In Goodman's quote, the elements are particular deductive inferences and rules of deduction. In a process of mutual adjustments, inferences and rules are revised ("ammended") in light of each other. If a state of coherence is reached, justification of all elements ensues.

But what exactly makes the coherentist solution virtuous? Coherence replaces narrow, or vicious circles with networks of interrelated elements, but is consistency and mutual support sufficient for justification? In this thesis, I attempt to offer an answer that takes the idea of a virtuous circle quite literally. RE involves *theoretical virtues*, that is, features of bodies of beliefs that are desirable from a broadly epistemic point of view. Theoretical virtues, e.g., consistency, simplicity or broad scope, originate from the appraisal of scientific theories, a topic which is extensively discussed in philosophy of science. For RE, theoretical virtues provide guidance during the process of mutual adjustments and they help to spell out coherence or systematisation in a state of equilibrium to go far beyond the basic ideas of consistency and mutual support.

My motivation to pursue this project is manifold. First, the "classic" and most widely discussed accounts of RE involve theoretical virtues implicitly, at best, and they do not assign them an active role in RE. Only more recent

and elaborate accounts of RE begin to render the role of theoretical virtues in RE more explicit and active. This trend culminates in the advent of a formal model of RE, which opens up unprecedented opportunities to operationalise theoretical virtues, and study their influence on RE computationally. Consequently, there is a lot of under-explored ground waiting to be investigated.

Next, the highly general discussion of RE left ideas vague, almost at the level of metaphors. The consideration of theoretical virtues in RE is no exception in this respect. There is general appreciation for theoretical virtues in philosophy of science, and there is a resemblance between RE and scientific inquiry. Does this suffice to motivate the involvement of theoretical virtues in RE? Or does this move threaten to import unsettled issues of theoretical virtues from philosophy of science to RE?

Finally, despite its great renown, objections to RE loom large. The early objections against Rawls have been raised again and again, even against updated accounts of RE. The controversy seems to come down to a deep-rooted disagreement about what RE can accomplish. The issue of these diverging impressions is aggravated by vagueness that besets the general discussion of RE. Critics suspect that RE might produce, and hence, justify completely different, and even absurd views. In contrast, proponents claim that RE has the means to prevent such outcomes. In the debate about RE, theoretical virtues do not take centre stage, and it seems to me that they often go unnoticed by critics. Thus, highlighting theoretical virtues may be a welcome addition to the defence of RE.

1.2 Goal and Methods

In view of the vast expanse of under-explored ground, we need to set a goal to guide the research. Otherwise, we might find ourselves wandering around aimlessly. The following research question shall guide this investigation:

> How can we integrate theoretical virtues into an account of RE such that they play an active role in addressing objections to the justificatory power of RE?

The question takes up the three motivating points from the previous section. First, I am interested in sharpening the profile for theoretical virtues in RE Next, my work is directed at overcoming the vagueness of RE in general, and in particular, addressing the issues of theoretical virtues that arise from

philosophy of science. Finally, I aim to achieve some progress in what I perceive as a stalemate in the debate about RE. I would like to move beyond mere plausibility considerations about whether RE lives up to its aspirations as an account of justification.

Note that the research question asks "How?", and accordingly, the answer, which I will give, is "Like this!". I aim for a proof of concept, illustrating that theoretical virtues can be rendered fruitful for RE. This answer does not aspire to be the only one, or even to be the best one.

I approach the research question with informal, formal and computational methods. First, RE and theoretical virtues are presented informally in the literature about them. In order to take up the ideas I apply the philosopher's usual tools such as literature surveys, conceptual work and the critical appraisal of views. Next, I rely on formalisation to overcome vagueness, and as a check for whether the ideas voiced in the informal literature can be spelled out thoroughly and consistently. Finally, a full-fledged formal model of RE provides the basis for a computational implementation, which allows to run simulations. Instead of relying on hunches whether RE might possibly yield desirable or detrimental results, we can analyse synthetic data generated by simulations on computers.

Preliminary Remarks At the outset of this project it is appropriate to remark on its general setting and how I will approach the tasks ahead.

I take RE to be general to philosophy. Most prominently, RE figures in ethics (see, e.g., Daniels, 1996; Rawls, 1971), but it is not restricted to ethical subject matters. RE also has a place in methodological discussions from logic (Goodman, 1983(1955); Peregrin and Svoboda, 2017) or rationality (Stein, 1996). Elgin (1996, 2017) bases her epistemology on RE, and still others take RE to be *the* method of philosophy in general (Keefe, 2000, Ch. 2; Lewis, 1983, x). Consequently, a methodological discussion of RE should not immediately force us to take sides in the great controversies of philosophy that RE touches upon. Hence, I will try to present my take on theoretical virtues in RE in an inclusive manner, and I will remain silent on many intriguing questions, for example whether RE, in the moral domain, is more compatible with realist or constructivist views.

Next, there is a helpful distinction introduced by Rechnitzer (2022, 11f) between RE as a *method* (a set of instructions), a *methodology* (theory or analysis of how we should proceed or spell out methods) or as an *epistemology* (a theory or analysis of what is epistemically valuable). The present project

moves between the understanding of RE as a methodology, in particular the methodological advise to include theoretical virtues, and RE as a method, especially the work required to render theoretical virtues fruitful as part of an applicable set of instructions. I attempt to remain silent on the most general level of RE as an epistemology. It seems to me, that the involvement of theoretical virtues in RE is compatible with different, high-level epistemological views about RE, e.g., as a purely coherentist (Tersman, 1993) or a weakly foundationalist theory of justification (e.g., Elgin and Van Cleve, 2014).

Finally, formalisation and modelling comes at the price of simplification and idealisation. I will work in a classical, propositional framework with deductive inferences. This is of course coarse-grained as it lacks, for example, representations for predicates or modal operators. Moreover, the framework cannot capture other forms of inferential relations, e.g., inductive, abductive or probabilistic reasoning. I will also assume that the reconstruction of parts of an agent's epistemic state as sets of sentences can be done successfully. In view of the aspired proof of concept, simple but workable foundations suffice as a starting point for future research.

1.3 Outline of Chapters

The work towards developing an answer to the research question is organised as follows: It is appropriate to decompose the research question into more manageable parts. To this purpose, the dissertation is split into three parts, and they roughly follow the methodological triad presented above. Part I is dedicated to arrive at an informal, but sufficiently elaborate understanding of theoretical virtues in RE. To begin with, I draw on elaborate accounts of RE to spell out the key ideas and central components of RE in Chapter 2. This serves to see how theoretical virtues are integrated into RE, and to provide a clear target for objections. In Chapter 3, I present objections to RE as well as prominent rejoinders. I focus on two objections that can be subsumed under the worry that RE is too weak as an account of justification, and hence put the justificatory power of RE into doubt: conservativity and no-convergence.

In Chapter 4, I survey the literature in philosophy of science, from which theoretical virtues originate, in order to get an idea of the roles and issues of theoretical virtues. In view of the issues of ambiguity and trade-offs, I propose that theoretical virtues for RE need to be configured, i.e., selected,

specified, weighted and aggregated, in view of the pragmatic-epistemic objectives pursued by RE.

Part II aims to overcome vagueness by formalisation. In Chapter 5, I take up the universal objective of coherence in RE, and I select and specify virtues to develop a substantive notion of coherence in a deductive framework. The second part of developing a configuration of theoretical virtues, aggregation and trade-offs, is addressed in Chapter 6. In Chapter 7, I introduce the full-fledged, formal model of Beisbart, Betz, and Brun (2021), which completely implements a configuration of theoretical virtues for RE. In Chapter 8, I analyse the formal model of RE. This lead to a series of technical, but highly interesting analytical results about weightings of theoretical virtues and other RE desiderata. This helps deepen our understanding of the inner workings of the model, and the results are useful to prepare, and later, interpret, computer simulations.

Part III serves to explore simulations on the basis the formal model, which is implemented as a computer program. I select a set of prospective parameters to be used in the subsequent simulation studies in Chapter 9. In Chapter 10, I operationalise three aspect of conservativity in the formal model, and examine whether the formal model performs better as the objection would lead us to expect. I proceed in a similar fashion for the study of convergence with RE on the basis of simulations in Chapter 11. Finally, in Chapter 12, I discuss the findings in the previous parts, see what lessons can be learned for the informal debate about RE, and provide an outlook to further research.

Part I

Reflective Equilibrium and Theoretical Virtues

Chapter 2

Reflective Equilibrium

In order to investigate the role of theoretical virtues in RE, we need an account of RE that goes beyond Goodman's (1983(1955)) description of RE as a "virtuous circle" or other classic expositions of RE (e.g., Rawls, 1971 ; Daniels, 1979). The goal of this chapter is to introduce a sufficiently elaborate account of RE that identifies key ideas and extracts components of RE from the particular contexts of their discussion or application in the literature, e.g., independent of Rawls' contractualist framework or Goodman's treatment of inductive inference. This also serves to provide a clear target for the objections in the next chapter.

The task of refining RE has been taken up many times before, and there is no need to start from scratch again. Consequently, I base the presentation of RE on recent and elaborate accounts of RE, such as those developed by, e.g., Elgin (1996, 2017), Baumberger and Brun (2017, 2021) or Rechnitzer (2022). As the elaborate accounts come with necessary complications of the originally simple ideas, I will attempt to develop a schematic illustration bit by bit to keep track of the increasing complexity of technical terminology and interrelationships between RE components.

The chapter is structured as follows: I present key ideas of RE in Section 2.1, and integrate them in an elaborate account of RE in Section 2.2. In Section 2.3, I collect various remarks on the relation between RE and justification.

2.1 Key Ideas of RE

Standard expositions of RE revolve around the central components of "judgement" and "principle" subsequent to Rawls' seminal work (1971). I will introduce the technical terms *commitment* and, respectively, *element of a theory* for these components shortly. For now, a pre-theoretical understanding suffices.

The main idea of RE goes as follows: An agent starts with their set of *initial* commitments about a subject matter. In an attempt to systematise the commitments, the agent comes up with elements of a theory that jointly account for the commitments. In a *process* of mutual adjustments, the agent revises the set of commitments and the theory in light of each other, striving to establish coherence among them. Commonly, such a *state* of coherence is called reflective equilibrium and it is supposed to provide justification to the theory as well as to the commitments. I suppose that we can identify the following key ideas of RE in this sketch:

i) a distinction between commitments and elements of a theory

ii) a state of equilibrium characterised by coherence among the set of commitments and the theory

iii) a process of equilibration, which starts from initial commitments about a subject matter, proceeding by mutual adjustments of theory and the set of commitments

Apart from the commonly drawn distinction between commitments and elements of a theory, above description makes an additional distinction, which is less sharp in the general literature about RE. We can distinguish between the *dynamic* aspect of an *equilibration process* of mutual adjustments, and the *static* aspect of a coherent *state of equilibrium*.

Let me illustrate how these key ideas are operative in an outline of RE by Scanlon (2003, 140–141):

> In broad outline (subject to further refinement) the method of reflective equilibrium proceeds in three stages. One begins by identifying a set of considered judgments about justice. These are judgments that seem clearly to be correct under conditions conducive to making good judgments of the relevant kind; that is, when one is fully informed about the matter in question, thinking carefully and clearly about it, and not subject to conflicts of interest or other factors that are likely to distort one's judgment. The second stage is to try to formulate principles that would "account for" these judgments. By this, Rawls means principles such that, had one simply been trying to apply them rather than trying to decide what seemed to be the case as far as justice is concerned, one would have been led to this same set of judgments.

> Since one's first attempt to come up with such principles is unlikely to be successful, there is a third stage in which one decides how to respond to the divergence between these principles and one's considered judgments. Should one give up the judgments that the principles fail to account for, or modify the principles, in order to achieve a better fit? It is likely that some accommodation of both of these kinds may be required. One is then to continue in this way, working back and forth between principles and judgments, until one reaches a set of principles and a set of judgments between which there is no conflict. This state is what Rawls calls reflective equilibrium.

Scanlon presents a three-staged structure of RE. First, there is Rawls' subject matter of justice, and a set of initial commitments ("judgements") about this subject matter. The commitments satisfy an additional demand of being "considered". The second and the third stage are part of a process of equilibration. In the second stage, one attempts to come up with a elements of a theory ("principles") that "account for" the commitments. Hence, we can observe a distinction between commitments and elements of a theory (i). Moreover, the initial commitments form the starting point for the RE process (iii). The third stage consist of mutual adjustments ("working back and forth") between commitments and elements of a theory. Unaccounted commitments may be given up or elements of a theory may be modified to achieve better fit between them (iii). Finally, a state of RE is reached if there are no conflicts between the set of commitments and the theory (ii). The coherence of this state consists of the absence of conflicts as well as the accounting-for-relation between the theory and the commitments.

Notably, theoretical virtues, or anything faintly reminiscent thereof, are absent from Scanlon's description of RE. This holds also for many other expositions of RE, for which we could read theoretical virtues into only with substantial interpretational effort. For example, the frequent use of "systematic principles" or that principles "systematise" judgements (e.g., Daniels, 1979) may hint at the involvement of theoretical virtues in RE. However, the role of theoretical virtues will become more apparent if we spell out RE in more detail.

2.2 Elaborating RE

In the following paragraphs, I take up the key ideas of RE and cast them into an informal account of RE that allows for further elaboration. Fortunately, I do not have to start from scratch, but I can draw on years of work towards the elaboration of RE. Two of the most detailed accounts can be found in (Rechnitzer, 2022) and (Baumberger and Brun, 2021). An attempt to condense the result of elaborating RE into a schematic overview is depicted in Figure 2.1.

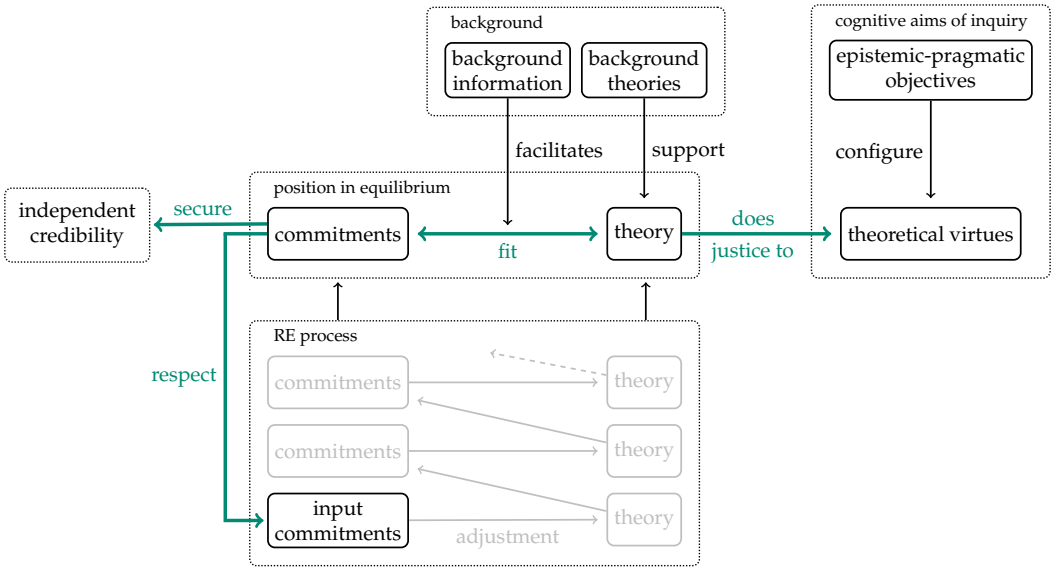


FIGURE 2.1: A diagram of RE components and demands on RE states and processes. Similar figures can be found in (Brun, 2020; Rechnitzer, 2022)

We go through the details of Figure 2.1 one by one, and spell them out. Apart from the key ideas, this includes desiderata and requirements that arise from the roles and the interplay of components. I use this this occasion to introduce and fix further terminology that I use throughout the project.

Commitments Commitments, or better sets thereof, are the first component that is present in literally all accounts of RE.[1] Commonly, commitments go under name of "judgments" (e.g., Rawls, 1971), but also "moral intuitions", or "beliefs" (Daniels, 1979) have taken up the same roles. I will follow Scheffler (1954), Elgin (1996) and Brun (2020) in using *commitment* as a technical term, escaping unwanted connotations that they must be in some sense explicit or conscious. Commitments comprise judgements, beliefs and the like as propositional attitudes with a commitment to a minimal epistemic status

[1] For the sake of having a more reader-friendly text, I will occasionally speak of "commitments" instead of "a set of commtiments".

(Baumberger and Brun, 2021, 7930). The modes of propositional attitudes of commitments, which I consider for this project, are accepting, rejecting or remaining silent (suspending). I will discuss the epistemic status of commitments later. Focusing on commitments as propositional attitudes excludes non-propositional elements from entering RE at this level, such as values or dispositions to act in certain ways. Still, it is my contention that propositional elements cover a wide range of philosophically interesting cases and allow for rigorous formalisation.

Commitments are frequently distinguished from elements of a theory, by the particularity of their contents. Commitments are taken to concern particular cases, but this need not be. Already Rawls indicated that agents have commitments "at all levels of generality" (1974, 8). Alternatively, we can distinguish commitments and elements of a theory by their function (Baumberger and Brun, 2017, 2021).

What are the functions of commitments? Foremost, commitments form the *input* and take up the intriguing idea that RE does not start from nowhere. They reflect the views an agent holds about a subject matter before, or independently of, engaging with RE. Next, commitments serve as a touchstone for theories, which are devised during RE to account for them. According to Baumberger and Brun (2021), commitments delineate a subject matter for an RE inquiry. Thus, commitments provide a point of reference for the question whether the subject matter has been changed too drastically during RE. As points of reference and as touchstones, commitments are relevant to the static aspect of equilibrium states as well as the dynamic aspect of equilibration processes. For the latter, the function of commitments as inputs is accentuated because they provide the point of departure.

Recent accounts of RE subdivide input commitments further into *initial* and *emerging* commitments (Baumberger and Brun, 2021; Rechnitzer, 2022). A set of initial commitments are given at the outset of RE forming the starting point for the process, emerging commitments may arise during equilibration due to newly available information. Emerging commitments are distinct from merely derived commitments, that can be inferred from the theory.

Besides input commitments, we can distinguish *current* commitments at a specific point in an equilibration process, and *resulting* or *output* commitments reached at the end of equilibration.

At this point, we do not demand that input commitments exhibit additional features that are desirable from an epistemic viewpoint. Thus, input commitments may as well include prejudices and biases. As a set, input

commitments may be inconsistent, fail to be closed under inferences, hang together loosely, and they may not be applicable to new cases. To put it in a nutshell, input commitments comprise our unsystematic views about a subject matter at the outset of RE, they form a "motley crew" (Elgin, 1996, 102). A goal of RE is to systematise the views at hand.

Elements of a Theory The second, universally present component in accounts of RE are the elements introduced to systematise commitments. Commonly, they are labelled "principles", but here, I rely on the technical term of *element of a theory* to denote a proposition about a subject matter. Sets of such elements are called *theories*. For the sake of simplicity, and symmetrical to the restriction of commitments, I do not consider non-propositional elements of theories, such as categories, diagrams and graphs (Baumberger and Brun, 2017, 3).

Again, although it runs against (Rawls, 1974, 8), the distinction between elements of a theory and commitments is commonly drawn according to their generality. Elements of a theory are supposed to be general principles rather than judgements about particular cases. As noted in the previous section, I prefer to distinguish them functionally. Elements of a theory serve to organise a subject matter in a more systematic way than commitments. In contrast to a mere collection of commitments, elements of a theory may reveal an underlying structure among commitments, or they may be applicable to new cases, which are not covered by the current commitments.

There are no restrictions to elements of a theory with respect to the generality of their content. Further demands on theories will arise from the specification of the metaphorical "equilibrium" between theories and commitments, as well as from epistemic goals that guide the effort of systematisation.

A pair of commitments and a theory is called a *position* (Rechnitzer, 2022, 18; Baumberger and Brun, 2021, 7926). It represents central parts in the epistemic state of an agent that engages with RE. Figure 2.2 displays a position, which is the central building block to RE in Figure 2.1.

FIGURE 2.2: A position, consisting of a set of commitments and a theory, is part of an agents' epistemic state.

Process of Equilibration The position represents parts of an agent's epistemic state, which undergoes changes in a dynamic equilibration process. This process does not start from nowhere, but from the input commitments about a subject matter. RE accounts come equipped with an idea of a process of "mutual adjustments", or "going back and forth" between commitments and elements of a theory aiming to establish a state of equilibrium.

In view of positions consisting of a set of commitments and a theory, there are two kinds of revisions: theory adjustments and commitment adjustments. The adjustment operations include the addition, removal, or modification (e.g., by negation or qualification) of commitments or elements of a theory. Figure 2.3 depicts an RE process schematically.

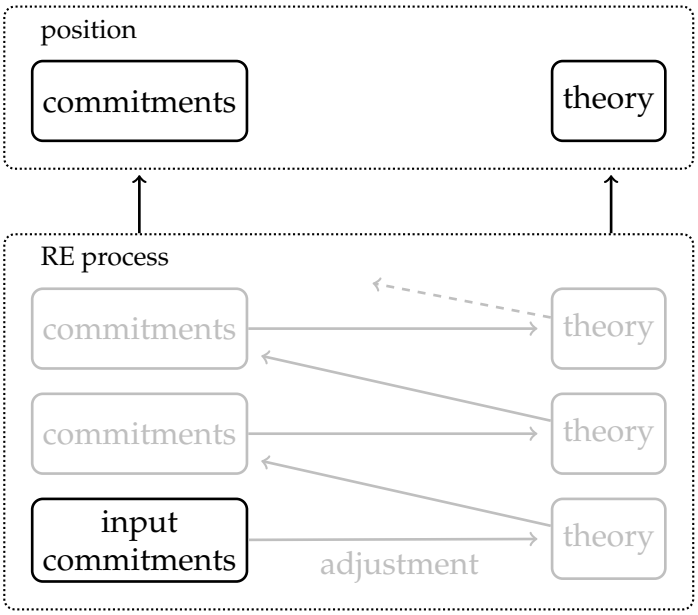


FIGURE 2.3: Starting from input commitments, a process of mutual adjustments leads into the current position.

State of Equilibrium (Incomplete) The aspired goal of the dynamic process of equilibration is to reach a state of equilibrium. The metaphorical notion of "equilibrium", and in its place notions like "agreement", "match", or "fit" are used to refer to a mostly unspecified state of coherence between commitments and the theory. What can we say about a position being coherent?

In its weakest characterisation, coherence coincides with consistency: Commitments and elements of a system are consistent with each other, if and only if their union does not contain or allows to infer a contradiction. The weak requirement of consistency is then supplemented with a more specific notion

of "hanging together", which is often expressed in terms of inferential relations between the set of commitments and the theory. Here, "inference" is broadly construed, that is it is not restricted to deductively valid arguments, it may as well include defeasible reasoning (Baumberger and Brun, 2021). Let *account* denote the positive inferential relation from the theory to commitments: Commitments are accounted for by the theory if the former are inferrable from the latter. In addition, let *fit* denote the relation between the commitments and the theory characterised by consistency and account.[2] Figure 2.4 complements a position with the internal relation of fit between the set of commitments and the theory.

FIGURE 2.4: The commitments and the theory of a position fit together if they are consistent with each other, and if the theory accounts for the commitments

In view of elaborate accounts of RE, however, fit still does not capture the idea of an RE state in its entirety. By now, we have covered the key ideas (i)–(iii) from Section 2.1, but there is much more to be found in elaborate accounts of RE. At his point, we have an incomplete account of so-called "narrow" RE, at best. It is "narrow" in view of extensions by additional components, and it is incomplete due to lack of further demands on components that arise from their roles and their interplay. Thus, it is important to continue elaborating RE in the following paragraphs.

Background The inclusion of a third component, *background theories*, has been made prominent by Daniels (1979). The idea is that a theory should not only account for the commitments, but that background theories should stand in support of the theory. Counteracting the flagrant circularity of fit between a set of commitments and a theory motivates the need to "widen the circle of justificatory beliefs" (Daniels, 1996, 1). As examples, in the context of Rawls' theory of justice, Daniels (1979, 260) suggests that "a theory of the person, a theory of procedural justice, general social theory, and a theory of the role of morality in society" serve as background theories.

[2] In contrast to (Baumberger and Brun, 2021; Goodman, 1983(1955); Rechnitzer, 2022), I refrain from using "agreement" between a set of commitments and a theory, so that we are able to speak about (dis-)agreeing positions in the course of dealing with questions about convergence in RE later.

Similar to the relation of account between the theory and the commitments of a position, which is in the foreground of an RE inquiry, the support relation between background and foreground theories is spelled out in terms of consistency and inference: A foreground theory should be consistent with or be inferrable from background theories (e.g., Daniels, 1979, 258).

What are the consequences of adding background theories to an elaborate account of RE? Positions become triples of a set of commitments, a theory and a set of background theories. Now, the aspired RE states should exhibit a three-way equilibrium. Moreover, every component of the triple is, at least in principle, revisable in view of the others.

However, this makes things significantly more complex, especially for equilibration processes. Rules or illustrative examples of adjustments of background theories are absent from the literature. Thus, I am sympathetic to the view expressed in elaborate accounts, that the background is "relatively fixed" (Rechnitzer, 2022, 32) or "treated as independently justified to some degree" (Baumberger and Brun, 2021, 7927). Note, that this does not render the background immune from revision. The distinction of foreground and background results from the epistemic project at hand, and for its justification, the background may come to the fore in another epistemic project pursued with RE. Adjusting the background theory (now in the foreground as part of a position) is then another process of RE.

There are more components to the background than theories. Background *information* may be needed to establish inferential relations between elements in the foreground (Baumberger and Brun, 2021, 7927). A commitment (e.g., a moral judgement about a particular case: "Tax fraud is morally wrong") may follow from a principle ("Lying is morally wrong.") only when there is relevant background information (e.g., that the principle is applicable to the particular case: "Tax fraud consists in lying to the government about your financial situation."). Rechnitzer (2022, 33) sees the need of background *assumptions* and *stipulations* in order to "get the process going" in applying RE.

Figure 2.5 presents the background as an additional element to RE.

Demands on Commitments Apart from introducing additional elements, one can also require that available components satisfy additional demands. As it stands, RE does not restrict (input) commitments, and hence the worry goes that intuitively absurd commitments could influence the process, or even worse, be part of the resulting position. Hence a "filter" may be set

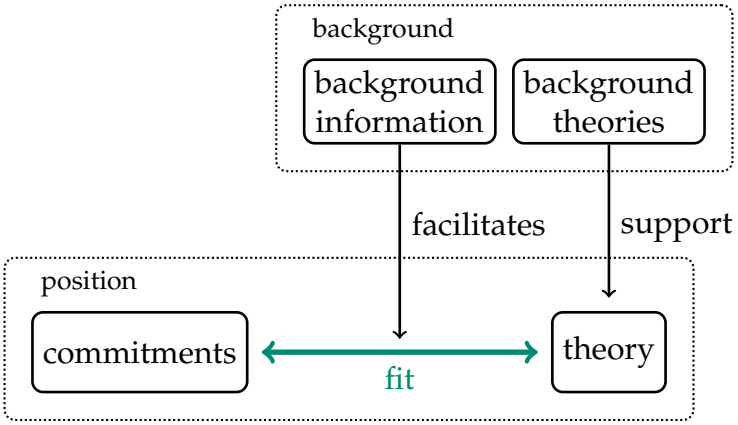


FIGURE 2.5: The background includes information that facilitates the fit between commitments and theory, and it provides theories that stand in support of the theory in the foreground.

up at the outset of RE to weed out what does not qualify as an input commitment.

The most prominent proposal in this respect is Rawls' additional demand of allowing only *considered* judgements as inputs to RE. According to Rawls, considered judgements are those made in conditions that are favourable to reason correctly (Rawls, 1971, 47–48). We do not hesitate to make those judgements and have confidence in them, they are not influenced by self-interest or arise from fear or when we are upset. However, Rawls' proposal to restrict input commitments to considered judgements attracted fierce criticism and is hotly debated to this day. For a discussion of considered judgements in view of objections to RE, see Section 3.4.

Another demand on input commitments went less noticed. Goodman (1952, 162–163) suggests to take a statement's *initial credibility* into account when we aim for a coherent system with a tie to fact. Initial credibility is a weak epistemic status. It is not certainty and it does not render statements immune from revision. Elgin takes up this idea and requires a "tether" to *initially tenable* commitments (Elgin, 1996, 101–107, 128; 2017, 64–65). Initially tenable commitments are our current best guesses about a subject matter, the sentences we have some, but not sufficient, reason to accept. Still, we need a reason to give up an initially tenable commitment. Note that Elgin's condition may be weaker than Rawls' filter: Elgin allows all sorts of commitments at the outset of RE, and lets equilibration weed out problematic commitments.

A variation of this theme stems from Baumberger and Brun (2021). They

distinguish *independent* and *initial* credibility, i.e., the credibility that a commitments has independent of coherence is not to be conflated with its credibility at the initial stage of an equilibration process (Baumberger and Brun, 2021, 7934). Emergent commitments may be independently credible even though they are not part of the initial commitments at the beginning of an RE process. Resulting commitments may not inherit independent credibility from initial or emergent commitments, as they may be given up or lose their credibility during the process. Consequently, Baumberger and Brun (2021, 7935) propose that securing some independent credibility for resulting commitments is a separate requirement for a position to be in a state of equilibrium, which is depicted in Figure 2.6.

FIGURE 2.6: The non-coherentist demand that a position secures some credibility independent of coherence considerations.

As mentioned before, it is a classic idea of RE that no component is immune from revision, especially the commitments (e.g., Daniels, 1979, 267). This bears the risk of diverting "too far" from the starting point of RE. If the commitments were merely hauled by systematic theories without limitations to commitment adjustments "there would be no guarantee that the process of equilibrating would not in fact change the subject" (Baumberger and Brun, 2021, 7932). This brings Baumberger and Brun to demand that adjusting the commitments should not change the subject too drastically in comparison to input commitments, which delineate the subject matter in the first place. Current commitments that arise from adjustments to the input commitments should *respect* the input. Respect does not amount to requiring a substantial overlap of input and current commitments, but that an agent is able to give reasons (Elgin, 2017, 64) or explain (Baumberger and Brun, 2021, 7932) why an input commitment was given up, replaced or adjusted in another way.

Note that the requirement to not change the subject by respecting input commitments is different from respecting the input. Respecting means that we need to have reasons to give up initial commitments, but this does also not guarantee that credibility is transferred from initial to resulting commitments.

There is no clear cut criterion for what counts as an overly drastic change of the subject, and the question needs to be addressed on basis of other elements from the context of an RE setting, i.e., the subject matter, the epistemic-pragmatic objective or background theories (Rechnitzer, 2022, 25; Baumberger and Brun, 2021, 7933). Figure 2.7 displays the "tie" of respecting the input commitments during an RE process. Note that the demands on commitments concern the commitment adjustment steps in a process of equilibration as well as the resulting state.

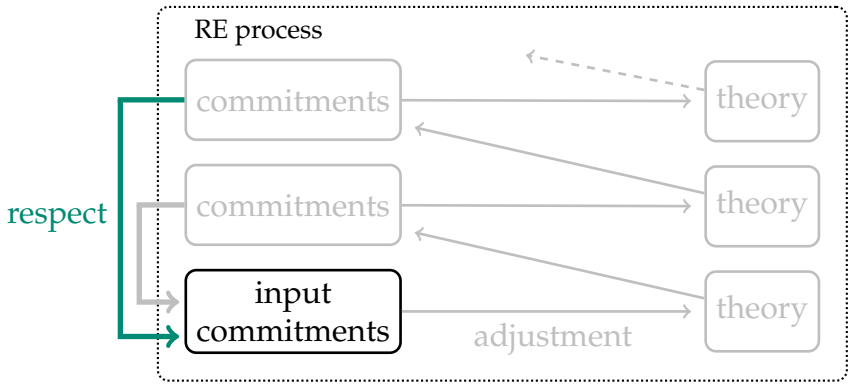


FIGURE 2.7: Current commitments should respect the input commitments meaning that one can make it plausible that commitment adjustments did not change the subject.

Demands on Theories At this point, the driving force behind the process, which guides revisions, stems from the mostly vague characterisation of equilibrium states mentioned so far. The adjustments of a set of commitments and a theory aim to render the components consistent with each other, striving for a theory that accounts for the commitments. However, this force may not be "driving" at all. Consistency is easily achievable by removing contradicting elements, and mutual support can be secured by repeating the list of commitments as theory. An agent might not be forced to revise their views, and they might just stick with their initial commitments.

Baumberger and Brun (2021, 7928) identify the often implicit allusion to epistemic goals as "the key driver of RE". It is a commonplace in the literature about RE that theories or principles should be systematic. Theories should systematise the commitments about a subject matter, but this demand was left mostly unspecified. Baumberger and Brun (2021) take a step towards spelling out systematicity in RE by demanding that theories *do justice* to epistemic goals. This is the entry point for theoretical virtues to RE. Epistemic

goals include theoretical virtues, which are prominently discussed in philosophy of science, e.g., accuracy, consistency, simplicity, broad scope, and fruitfulness (Kuhn, 1977).

Theoretical virtues pull in different directions, making trade-offs unavoidable. For example, increasing the simplicity of a theory may come at the cost of narrow scope. This motivates the use of "doing justice" instead of "realising" these goals (Baumberger and Brun, 2017, 178), as they may not be satisfiable all at once or not to a maximal degree. Competing epistemic goals need to be configured in view of epistemic-pragmatic objectives of an RE inquiry. For example, Rechnitzer (2022, 115) selects practicability, determinacy, broad scope, and simplicity in view of the pragmatic-epistemic objective of "justifying an action-guiding moral principle that is applicable to the subject matter of precaution and precautionary decision-making" (Rechnitzer, 2022, 101).

Figure 2.8 illustrates the additional demand of doing justice to epistemic goals schematically. Note that this demand on theories concerns the theory adjustment steps during the process of equilibration as well as the state of equilibrium.

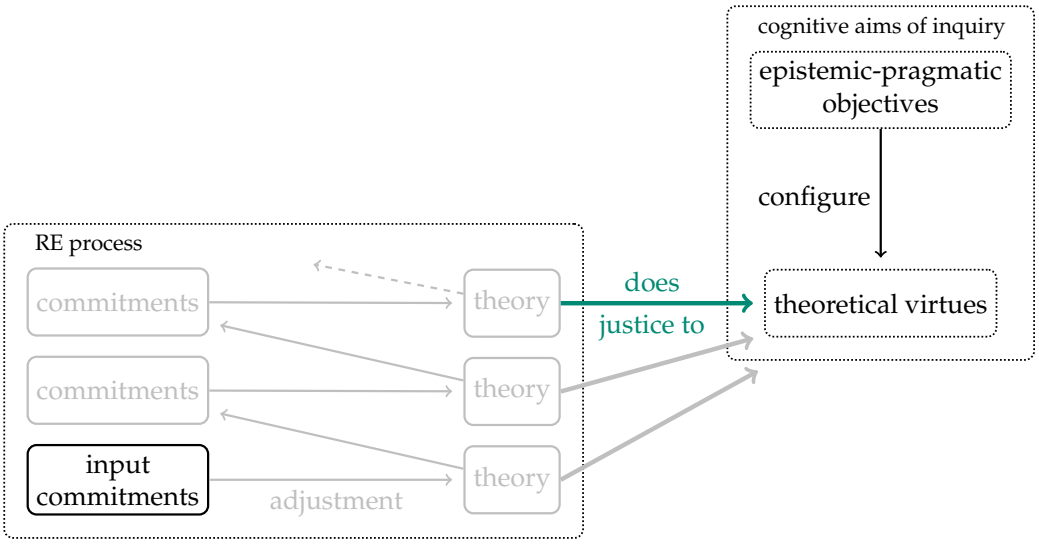


FIGURE 2.8: The driving force of theory adjustments in a process of equilibration is systematisation. A theory is systematic insofar it does justice to epistemic goals, especially theoretical virtues. Epistemic goals are configured by the epistemic-pragmatic objectives of an RE inquiry.

2.2.1 Summary: State of Reflective Equilibrium

Now, that all elements of Figure 2.1 have been introduced, we are in a position to complement the hitherto incomplete idea of a state of RE. Demands exert force on the position, pulling in different directions and guiding a process of equilibration (Baumberger and Brun, 2021, 7935).

First, fit is an internal force of attraction between the commitments and the theory of a position, and it is mediated by the background. The commitments and the theory should be consistent with each other and the theory should account for the commitments. However, merely achieving fit leads to an incomplete characterisation of a state of equilibrium. Second, there is a "progressive" pull of doing justice to epistemic goals aiming to render the theory systematic. Third, there is a "conservative" pull on the side of current commitments towards the input commitments so that the subject is not changed. In addition, commitments should have some credibility that is independent of RE considerations, and background theories can stand in support of a position's theory in the foreground.

Fitting commitments and theory to each other, doing justice to epistemic goals, respecting the input, and securing independent credibility stand in need of balancing as well. If a balance between all "forces" can be struck, a state of equilibrium is reached. Thus, a position, consisting of a set of commitments and a theory, is in a *state* of equilibrium if it satisfies the following conditions.[3]

1. The set of commitments, the theory and the background are in equilibrium characterised by consistency, account, and support mediated by inferential relations among them.

2. The theory does justice to a configuration of epistemic goals, especially theoretical virtues.

3. The current commitments respect the input commitments by not changing the subject, and they secure independent credibility to some extent.

A fourth condition arises from the need to balance the forces. It has been introduced by Elgin (2017, 66, 87f; 1996, 107) and taken up by Rechnitzer (2022, 35):

4. The position is at least as reasonable as any available alternative in light of the initial commitments.

[3]The conditions could be split up further (for a similar list, see (Rechnitzer, 2022, 35), but in this form, they are grouped according to the components the conditions apply to.

These conditions go beyond standard descriptions of RE in the literature. Commonly, the relation of fit is featured prominently as characterisation of "equilibrium" in descriptions of RE, while respecting input commitments, securing independent credibility, and doing justice to epistemic goals are conveyed implicitly, at best. Still, Figure 2.1 does not depict everything that there is to elaborate RE. For example, the distinction of initial and emergent commitments is missing. Moreover, the idea that the balancing of forces, or more specifically, the configuring of theoretical virtues may be developed during equilibration and undergo change in feedback loops is hardly representable in such a picture.

2.3 RE and Justification

Elaborating the components and demands for RE states and processes helps to overcome the vagueness of describing the state of equilibrium as "everything fitting into one coherent view" and the process of equilibration as "working back and forth". Another aspect, which is rarely at the centre of attention, is the relation of RE and justification. It is obvious from the literature that RE is supposed to justify views, but the discussion hardly ever focuses on the exact relation between RE and justification.[4] This is troubling, as it adds to the vagueness of already hazy RE accounts, and it poses the threat of talking past each other. In the following paragraphs, I present various aspects of how RE may be related to justification, on which even proponents of RE tend to disagree.

I cannot to settle these issues neither at the outset of this project nor at the end, as an in-depth treatment goes beyond the scope of this work. However, the consideration of theoretical virtues in RE may provoke new answers or push us to one side in the controversy rather than another.

There are also aspects of RE and justification that I cannot cover, and for which I do not see that theoretical virtues would play a decisive role. For example, "being justified" and "being true" can come apart in RE as an account of justification. For a discussion of the relation between justification and truth in RE, see (Tersman, 1993, 94–114). The epistemicity of theoretical virtues, narrowly understood as their truth-conduciveness, is highly disputed in philosophy of science (see Chapter 4), and they figure in realist, as well as anti-realist accounts. Hence, including theoretical virtues in RE does

[4] A notable exception is Hahn (2000, A3.3) who discusses the analysis of justification in terms of RE as necessary and sufficient condition.

does not require nor entail realism about its domain of inquiry (e.g., morality).

Coherentism and Foundationalism On many occasion, RE is presented as a *coherentist* account of justification. Take for example Rawls' classic exposition, in which he distances himself from the foundationalist aspirations of intuitionism:

> A conception of justice cannot be deduced from self-evident premises or conditions on principles; instead, its justification is a matter of the mutual support of many considerations, of everything fitting together into one coherent view. (Rawls, 1971, 21)

Clearly, coherence plays an important role in RE, but the crucial question is whether RE is *purely* coherentist or not. For a discussion of this question, see (Ebertz, 1993). Tersman (1993, 2018) promotes an account of RE that is based exclusively on coherence.[5]

The presentation of the elaborate account of RE from above suggests that RE may not qualify as a purely coherentist account of justification. Note that this can be seen an asset as it dissolves the worry that RE creates justification *ex nihilo*. The tie to a minimal epistemic status of commitments that is independent of coherence considerations, and the demand to respect input commitments, may locate RE somewhere between foundationalism and coherentism. There is a helpful distinction of strengths of foundationalism developed by BonJour (1985, 26–28):

(Strong Foundationalism) There are basic beliefs are logically infallible, and thus unrevisable.

(Modest Foundationalism) There are basic beliefs are justified non-inferentially to a degree that would qualify them as knowledge if the beliefs happen to be true.

(Weak Foundationalism) There are basic beliefs are justified non-inferentially to a relative low degree that is not sufficient for knowledge.

Rawls' quote from above suggests that he rejects strong foundationalism. Moreover, it might be more promising to construe RE as weak rather than

[5]Note that a purely coherentist interpretation of RE does not need to give up on allegedly non-coherentist RE elements, such as independent credibility, by including second-order beliefs about the reliability of (first-order) beliefs (Tersman, 2008, 399).

moderate foundationalism. Otherwise, RE relies on independently *justified* input commitments, which attracted criticism (see Section 3.4). A weakly foundationalist interpretations of RE are promoted by, e.g., Rechnitzer (2022), Elgin and Van Cleve (2014) or Brun (2014). Note that weak foundationalism does not render coherence irrelevant as an epistemic goal of RE. Coherence in the output may boost the weak epistemic status of the input above the threshold of what counts as justified. Moreover, formal results of Hansson (2007) suggest that weakly coherentist and weakly foundationalist requirements are compatible with each other.

Finally, there are also epistemic goals for theories, and in particular theoretical virtues, that are not related to coherence, e.g., precision, conceptual clarity and visualisability (Baumberger and Brun, 2021, 7931), or practicability and determinacy (Rechnitzer, 2022, 118). Thus, doing justice to such epistemic goals in RE may also add a further, not purely coherentist, spin to RE.

Consequentialist and Proceduralist Justification RE comes with a distinction between a process of equilibration and a state of equilibrium. So, one can ask whether and how process and state contribute separately to justification. At first sight, the state of RE relates to justification in a straightforward manner. A state of equilibrium exhibits features that are desirable for justification from a coherentist or a weakly foundationalist point of view. From a consequentialist stance towards RE, being in an state of equilibrium is all that is needed to be justified, or to use Goodman's words:

> The process of justification is the delicate one of making mutual adjustments between rules and accepted inferences; and in the agreement achieved lies the *only* justification needed for either. (Goodman, 1983(1955), 64, emphasis added)

So we can read Goodman to put emphasis on agreement (the state of equilibrium) only for justification, and similarly, there are no hints of proceduralist aspects of justification in Rawls often-cited remark that justification by RE is a "matter of the mutual support of many considerations, of everything fitting together into one coherent view" (Rawls, 1999a, 19). From a consequentialist perspective, the process has merely instrumental value insofar it brings forth the coherent state of equilibrium that exhibits justificatory power.

In contrast, Scanlon (2014, 79–84) emphatically endorses a purely proceduralist account of RE, when he writes:

> The justificatory force, if any, of being among the beliefs we have arrived at in reflective equilibrium must lie in the details of how the equilibrium is reached. (Scanlon, 2014, 79)

For a critical assessment of two line of arguments that speak in favour of proceduralism, see (Baumberger and Brun, 2021).

Occasionally, the proceduralist stance is refined to an *imperfect procedural epistemology* (Elgin, 1996, 2017; Rechnitzer, 2022). As it stands, the process of equilibration is an *imperfect* procedure, i.e., it does not guarantee that its outputs meet a process-independent "criterion of correctness" (Elgin, 1996, 4). Note that the conditions for a state to be in equilibrium refers back to the initial commitments, but it does not require that one has reached the state by going through an actual process. Hence, the conditions provide the process-independent criterion. Even though the process of equilibration aims for a position to be in a state of equilibrium, nothing in the process guarantees that this goal is achieved.

Application and Reconstruction The fact that RE accounts are equipped with an equilibrating process may have cultivated the view that RE is best understood as a *method* of justification. This may suggest that RE has a descriptive component how an agent actually reaches a state of RE, or a prescriptive component as a kind of "recipe", a set of instructions that can be followed to justify one's views.

This idea is underwritten by more or less rigorous attempts to apply RE in various contexts and with different aims in mind. For an overview of such applications, and the most detailed case study to this day, see (Rechnitzer, 2022).

Already Goodman (1983(1955), 65) distances himself from the descriptive and prescriptive approach to apply RE, and many followed (e.g., DePaul, 2006, 599; Keefe, 2000, 42). Hence, recent and elaborate accounts opt for a reconstructive approach to RE (Baumberger and Brun, 2021; Elgin, 2017).[6] According to this view, a position is justified if it can be reconstructed as a position in a state of RE resulting from an RE process irrespective of whether an agent actually went through such a process.

[6] Even Rechnitzer (2022, 238) sees her detailed presentation of a process at book length as a "cleaned up" reconstruction.

Personal and Doxastic Justification RE is proposed or criticised as an account of justification. But what exactly is justified by RE? Scanlon brings up the following distinction:

> "Justification" can be understood in two ways. On the one hand, to claim that a *principle* or *judgment* is justified is to say that it is supported by good and sufficient reasons. But we also speak of a *person's* being justified in holding a certain view. To claim that he is is to claim that he holds that view for reasons that he reasonably takes to be good and sufficient. A person can be justified, in this sense, in accepting a principle (for certain reasons) even though the principle itself is not justified because, say, there are other factors (which he could not be expected to be aware of) that undermine the justificatory force of the considerations he takes to be reasons for it. [...] In the case of reflective equilibrium, however, it may not be clear which sense of justification is involved. A person may be justified in accepting a principle if it accounts for his or her considered judgments in reflective equilibrium and the person has no reason to modify or abandon these judgments. But it does not follow that this principle is justified." (Scanlon, 2003, 140)

On the one hand, we can take the *position*, consisting of a set of commitments and a theory, to be the object of justification by RE. It is important to keep this in mind as work in the literature sometimes focuses on only one outcome of RE. Daniels (1979), for example, stresses theory acceptance as outcome of RE. From the level of positions, justification "trickles down". A theory is justified if it belongs to a position in RE, and an element of a theory is justified as a part thereof. Analogously, an individual commitment is justified if it is an element of a justified set of commitments, which belongs to a position in RE. On the other hand, an *agent* may be justified in accepting the elements of a position in RE.

So, there is a distinction of general epistemology at play between justification that pertains to a cognitive agent having an attitude towards a proposition (*personal* justification), or to the proposition itself (*doxastic* justification).[7] The distinction becomes apparent by considering the scope of "justified" in the following formulations for an agent S and a proposition p:

(Personal Justification) S is justified in believing that p.

[7] I rely on the terminology that is used by Engel (1992) and Koppelberg (2012).

(Doxastic Justification) *S*'s belief that *p* is justified.

The distinction is commonly drawn as follows. Engel (1992, 138), for example, ties doxastic justification to a sufficiently high objective probability of being true. Doxastic justification, in turn, can be spelled out in deontological terms (epistemic responsibility or blamelesness), or by means of the exercise of intellectual virtues (Koppelberg, 2012, 313).

It is important to note that personal and doxastic justification can come apart. The conflation (or explicitly assuming the equivalence) of personal and doxastic justification arguably caused that internalists and externalists talked past each other in their debate about justification (Engel, 1992). In order to avoid confusion, the distinction should be transferred to the debate about RE, which, at this date, is still characterised by a flip-flopping usage of doxastic and personal justification in the literature on RE. Strong (2010, 127), for example, portrays (Rawls, 1999a, 44) as being inclined towards personal justification. In contrast, Daniels (1979, 257, footnote) discards personal justification and aims to address theory acceptance, which seems to be related to doxastic justification.

Introducing the distinction does not resolve the quarrels about RE as a method of justification, but it is worthwhile to clarify what kind of justification is at stake when addressing the objections to RE. For example, the alleged lack of convergence in RE may not be that damaging as an objection to RE, if RE is understood as an account of personal justification.

Pluralism and Relativism The process of equilibration sets out from initial commitments, and a state of equilibrium is evaluated in light of them. Hence, the results of RE are already relativised with respect to the initial commitments. The elaborate account of RE reveals that RE involves many components and demands. To name a few, an agent engaging with RE needs to rely on a subject matter, background theories and background information, epistemic-pragmatic objectives that configure theoretical virtues, and sources for independent credibility. Consequently, the process of equilibration and its result, or the state of equilibrium depend on much more than just the set of initial commitments about a subject matter. I suggest to summarise the components, which are relevant to carry out an equilibration or to evaluate a state, as parts of the *epistemic situation* of an agent. The literature on RE provides discussions of relativisation to "epistemic circumstances" (Elgin, 1996, 143ff), or to "epistemological situation" (Scanlon, 2014, 80ff). Note that this also helps to highlight the provisional status of justification by RE.

If the epistemic situation of an agent changes, a state of equilibrium may be disrupted leading to further adjustments of a position.

Related to the relativisation of RE to the epistemic situation, the question arises whether justification by RE is pluralistic. There is a longstanding tradition of affirmative answers by proponents of RE (Goodman, 1983(1955), 63; Elgin, 1996, 135ff; Brun, 2022, 25; Rechnitzer, 2022, 20f). Even in view of identical epistemic situations, agents may reach different outcomes due to multiple available alternatives that strike equally good balances. In this case, both outcomes may be seen as justified, and pluralism ensues. Whether RE is overly pluralistic or relativistic is the object of fierce debate.

Analysis and Explication We can construe the conceptual relation between "RE" and "justification" in various ways. Two apparent approaches are analysis or explication. On the one hand, the relation may be analysed in terms of necessary and sufficient conditions (Hahn, 2000, 139–140):

i) A position is justified if and only if it is in RE. (RE is necessary and sufficient for justification)

ii) If a position is in RE then it is justified, but not vice versa. (RE is sufficient but not necessary for justification)

iii) If a position is justified then it is in RE, but not vice versa. (RE is necessary but not sufficient for justification)

iv) Being in RE is neither necessary nor sufficient for being justified.

Hahn (2000, 18–19) notes that i), ii) and iv) are brought up in the debate around RE. However, she comes to the conclusion that a discussion of how RE and justification relate requires an elaborate account of RE that goes beyond the metaphor that was introduced by Goodman and Rawls (Hahn, 2000, 140).

i) amounts to a successful conceptual analysis of the analysandum "justification". Here, the analysans "RE" is necessary and sufficient for justification. In view of the multiplicity of conditions for a state of equilibrium from above, we can speak of individually necessary and jointly sufficient conditions for justification. In the next chapter, I will focus on objections that put ii) into doubt.

On the other hand, "RE" can take the role of an explicatum in a Carnapian explication of the explanandum "justification". Here, "RE" is taken to replace "justification". To be in RE is what it means to be justified. Here, RE

explicates the pre-theoretical concept of justification by spelling out explicit rules for its use in a "well-constructed system of concepts" (Carnap, 1950, 3). In our case, the four conditions to be in a state of RE from above provide such rules. In contrast to necessary and sufficient conditions in analysis, the *adequacy* of an explicatum has to be evaluated in terms of its similarity to the explicatum, its exactness, fruitfulness, and simplicity (Carnap, 1950, 5). In this understanding, objections to RE put into doubt whether RE is sufficiently similar to the pre-theoretical concept of justification or adequate in other respects.

Chapter 3

Objections to Reflective Equilibrium

3.1 Is Reflective Equilibrium Too Weak?

A multitude of objections to RE as an account of justification has been raised over the years. The objections are intertwined, and it is difficult to completely disentangle all criticism. For the scope of the project at hand, I will focus on the strand of criticism that stems from the general worry that RE is *too weak* as an account of justification: conservativity and no-convergence. They stand out as longstanding and prominent issues, which are recognised by many proponents of RE. As the objections are directed at the weakness of RE, they will later provide a good target to examine whether theoretical virtues can strengthen RE as an account of justification.

Even though the objections are raised against vaguely characterised ideas of RE, I rely on the more elaborate ideas and the terminology developed in the previous chapter. This helps to render the objection as precise and strong as possible before we discuss the prospects and limitations of addressing them on the basis of elaborate RE, especially if theoretical virtues are involved.

Explicit phrasings of the alleged weakness of RE can be found from early critics of Rawls to more recent criticism (Lyons, 1975, 147; Little, 1984, 373; Kelly and McGrath, 2010, 326), and implicit allusion are even more abundant. How can we understand this concern of weakness? In its most general form, and expressed as a conceptual relation, we may suppose that some critics claim that

(Weakness) RE is not sufficient for justification,

and take this to be a serious flaw of a good (or even the best) account of justification. (Weakness) states that a set of commitments and a theory in a state of RE does not suffice to render this pair justified.[1]

One can argue for (Weakness) by making it plausible that the conditions for a state of RE obtain for a set of commitments and a theory, while they still lack epistemically desirable features related to justification, e.g., truth, reasonableness or objectivity. The following two strategies to object against RE can be seen to spell out and lend support to (Weakness).

(Conservativity) RE does not provide enough incentive for a substantial revision of initial commitments.

(No-Convergence) RE is not able to achieve converging, non-substantially-disagreeing outputs.

In this chapter, I aim to spell out (Conservativity) and (No-Convergence). The work towards this goal is structured as follows. (Conservativity) and (No-Convergence) are rendered more precise in Section 3.2 and Section 3.3, respectively. I discuss previous replies as well as the prospects and limitations of addressing the objection on the basis of theoretical virtues in Section 3.4.

The focus on (Conservativity) and (No-Convergence) under the umbrella of (Weakness) is just a small selection of objections. How do they interrelate and how do they relate to other prominent objections against RE? Answering this question may be useful to examine the prospects of addressing these objections with theoretical virtues as well. I will point to interrelationships among objections en passant. Note that there are also objections to RE that are not related to theoretical virtues. For example, Cummins (1998) argues that intuitions, which serve as initial commitments, are unreliable (see also Hare, 1973; Singer, 1974). However, Brun (2014) argues against the view that RE essentially involves intuitions. Other objections are directed at more general issues of coherentism. Setiya (2012), for example, argues that in face of disagreement, RE (understood as pure coherentism) has to rely on epistemic

[1] The formulation of (Weakness) suggests to read the relation between RE and justification as analysis (see Chapter 2.3). However, it also applies to RE as an explication of justification. It is problematic for both approaches if there are positions that are in RE but not justified. The objections cast doubt about the validity of the unqualified inference from RE to justification in the first approach. In the second, explicative approach, RE is *inadequate* according to the weakness objections, as it fails to be sufficiently *similar* to the pre-theoretical concept of justification. The worry goes that too many, or relevant pre-theoretically unjustified positions can be in a state of RE, and thus could be justified according to the explication with RE.

egoism to avoid scepticism. Finally, there is a specific objection against the truth-conduciveness of theoretical virtues in RE in (Kappel, 2006), which I present in Section 4.2, after surveying the treatment of theoretical virtues in philosophy of science in the next chapter.

3.2 Conservativity

I take the following passage of Harman (1986, 32) to be an interesting entry point to the discussion of conservativity from the viewpoint of RE, as he relates coherence and conservativity:

> The coherence theory supposes one's present beliefs are justified just as they are in the absence of special reasons to change them, where changes are allowed only to the extent that they yield sufficient increases in coherence. [...]
>
> According to the coherence theory, if one's beliefs are incoherent in some way, because of outright inconsistency or simple *ad hocness*, then one should try to make minimal changes in those beliefs in order to eliminate the incoherence. [...]
>
> It is important that coherence competes with conservatism. It is as if there were two aims or tendencies of reasoned revision, to maximize coherence and to minimize change. Both tendencies are important. Without conservatism a person would be led to reduce his or her beliefs to the single Parmenidean thought that all is one. Without the tendency toward coherence we would have what Peirce (1877) called the method of tenacity, in which one holds to one's initial convictions no matter what evidence may accumulate against them.

This passage brings up important ideas such as that the absence of incoherence does not motivate revisions, and the strive to increase coherence with minimal change.

It is important to know that conservatism, as a methodological principle, is also being discussed more generally in epistemology and philosophy of science.[2] It is an integral part to operations of belief change in the field of Belief Revision Theory (Alchourrón, Gärdenfors, and Makinson, 1985). Next, it plays a positive role as a virtue in the evaluation of scientific hypotheses

[2]For an overview in epistemology and a critical stance, see (Christensen, 1994).

(Quine and Ullian, 1978, 66f), and "our natural tendency to disturb the total system as little as possible" (Quine, 1951, 41) is involved in adjustments of a system of statements in view of recalcitrant data: "Conservatism figures in such choices, and so does the quest for simplicity." (Quine, 1951, 43). It seems to me that Daniels (1979, 262) alludes to this last quotation of Quine and transfers it to RE:

> [...] as in science, judgments about the plausibility and acceptability of various claims are the complex result of the whole system of interconnected theories already found acceptable. My guess -— I cannot undertake to confirm it here -— is that the type of coherence constraint that operates in the moral and nonmoral cases functions to produce many similarities: we should find methodological conservatism in both; we will find that"simplicity" judgments in both really depend on determining how little we have to change in the interconnected background theories already accepted [...]

The ideas at the nexus of coherence and conservativity surface in the debate about RE, and they inspire objections. So, where can we locate (Conservativity) in the literature about RE?

To begin with, I rule out a source of conservativity, which arises from the suspicion that RE involves elements that are unrevisable. Early critics of Rawls took considered judgements to "remain as fixed points" (Singer, 1974, 516) after their initial consideration, signifying that some initial commitments were unrevisable. Arguably, this is a misunderstanding if we take considered judgements to be initial as well as resulting commitments of RE. As inputs and during the process of adjustment, considered judgements are revisable, and treated "provisionally as fixed points" (Rawls, 1971, 18) for adjustments of a theory, but they are not "in principle immune to revision"(Rawls, 1999b, 289). As outputs, they do not need to be revised at that very moment, because they are in a state of equilibrium.

Singer (2005, 347) still suspects that conservativity is built into "narrow" RE:

> On this view the acceptability of a moral theory is not determined by the internal coherence and plausibility of the theory itself, but, to a significant extent, by its agreement with those of our prior moral judgments that we are unwilling to revise or abandon.
> (Singer, 2005, 345)

But even for a willing agent, RE may provide not enough incentive to revise or abandon their initial commitments. Some authors present (Conservativity) in relation to the inherent circularity of RE (Lyons, 1975, 146; Haslett, 1987, 307, Arras, 2007, 65). Circularity arises from the relation of "mutual support" in RE states given by the symmetric relation of fit between a set of commitments and a theory. However, fit seems to be easily reachable by minor adjustments, so that an agent "accomplished little more than a systematisation of his initial perspective (Haslett, 1987, 307), making RE a "technique for systematising and organising one's antecedent moral convictions" (Little, 1984, 373).[3]

Another presentation can be found in (Brandt, 1985). He presents RE as "sophisticated intuitionism" (Brandt, 1985, 6f) including a set of basic (underived), considered epistemic statements of any level of generality. The statements are held with different degrees of credence between 0 and 1. A system of statements is made coherent with adjustments following Goodman's procedure of mutual adjustments (Goodman, 1983(1955), 64), or Scheffler's idea of maximising credibility (Scheffler, 1954, 183). Brandt (1985, 7) objects "that the procedure seems to amount to no more than a re-shuffling of one's initial prejudices". He traces the re-shuffling back to a lack of evaluating the commitments independent of their coherence with each other (Brandt, 1985, 8).

More recent accounts also raise objections to RE on the basis of the suspicion that RE is overly conservative (de Maagt, 2017; Dutilh Novaes, 2020; Kelly and McGrath, 2010; Strong, 2010).

I take the following to be the gist of objecting to RE on the basis of (Conservativity). It targets the static aspect of equilibrium states, as well as the dynamic aspect of equilibration. The objection takes off from a weak characterisation of equilibrium states in terms of consistency or coherence. This may set the bar so low for what counts as a state of RE, that they are easily reachable by a mere streamlining of one's initial views with minor adjustments.

We can spell out the idea of a streamlining procedure further. A process of equilibration aspires to reach a state of equilibrium, but the latter is oftentimes vaguely characterised in terms of coherence, narrowly understood as fit between a set of commitments and a theory. In turn, fit has been spelled out in the previous chapter to consist of consistency and account. In this

[3]Note that those are examples of explicitly mentioning systematisation in RE, but without spelling it out to play a substantive role.

simplistic depiction of RE, incoherence is the key driver behind adjustments in an equilibration process. However, both consistency and account can or should be established with minimal adjustments (c.f. Harman's quote on page 33).

Even though the literature is mostly silent about it, it seems to me that classical propositional logic provides the intuitive backdrop for illustrating the weakness of coherence as consistency and fit. Classically, the logical consequence relation is monotone: If a proposition p logically follows from a set of propositions A, then p follows from any superset $B \supseteq A$.[4] In this view, an inconsistent set of propositions, i.e., a set which has some proposition p as well as its negation $\neg p$ among its consequences, cannot be made consistent by adding more elements. Thus, the strategy to resolve an inconsistency is to remove (as few as possible) elements resulting in a subset of the original set. Such a maximally consistent subset of an inconsistent set of commitments is conservative as it preserves as many initial commitments as possible without introducing new commitments.[5]

Consistency is the absence of contradictory consequences, and hence, it is worthwhile to look for more positively characterised features that have to be present for coherence to obtain. The most common positive ingredient to coherence, includes a support relation (e.g., given by inferences) among the involved elements in addition to consistency. The demand that the theory should account for the commitments implements this in the elaborate account of RE. Again, deductive inference seems to provide an intuitive approximation. A commitments is accounted for by a theory if it can be deduced from the theory (perhaps together with additional information or assumptions from the background). As noted by Baumberger and Brun (2017, 177) or Rechnitzer (2022, 22), this allows to follow a strategy to enlist all initial commitments as elements of a theory. Trivially, any element p accounts for itself in the weak sense that it can be inferred from itself ($p \rightarrow p$ is a tautology). This strategy is conservative as it does not require to revise commitments during subsequent adjustment steps. Clearly, such a strategy will fail to yield systematic theories which will bring us to theoretical virtues shortly. For now, we stick a little longer with the simplistic depiction of coherence in RE as mere fit between a set of commitments and a theory to ponder over

[4] This does not hold in non-monotonic logics deployed for, e.g., defeasible reasoning.

[5] The focus on maximally consistent subset of inconsistent sets of beliefs permeates the field of Belief Revision Theory (BRT). According to Rott (2000), BRT took up Quine's principle of minimal mutilation under the label of "informational economy", and made it to one of its dogmatic cornerstones.

(Conservativity).

So far, the strategies can be brought together in a *streamlining procedure*, which is schematically depicted in Figure 3.1. A streamlining of initial commitments could be achieved with the following instructions: If the initial commitments are consistent, choose a minimal axiomatic base in the commitments, i.e., a minimal subset of the commitments entailing all commitments as current theory. Otherwise, the initial commitments are inconsistent, and there is no axiomatic base. In this case, choose a maximally consistent subset of the initial commitments as well as a minimal axiomatic base for them. Adapt the commitments to the consequences of the current theory. Stop.

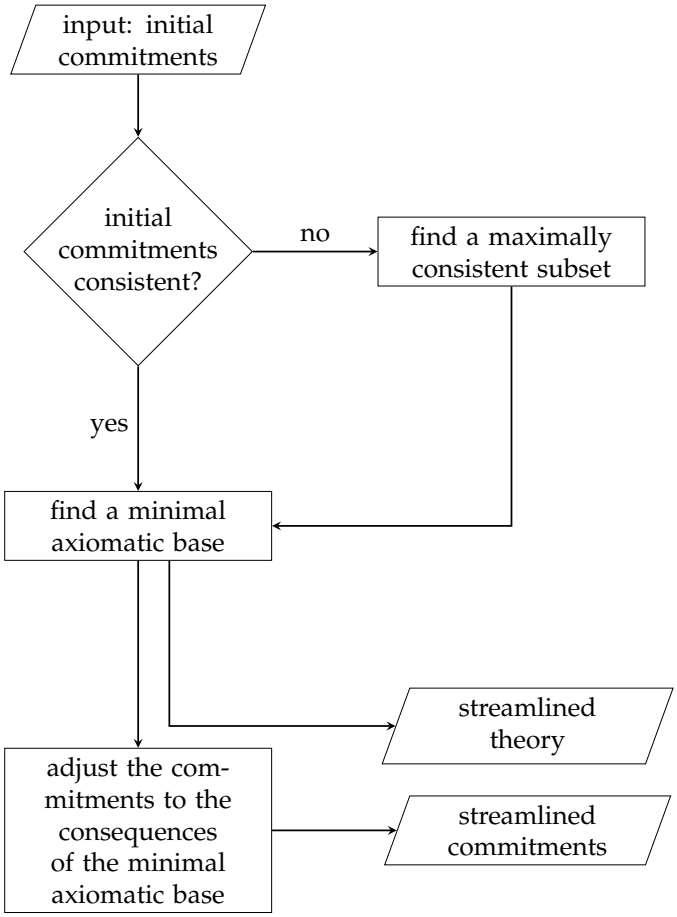


FIGURE 3.1: A procedure that streamlines the input with minimal effort.

The streamlining procedure is conservative by design. Consistency is established (if necessary) by maximally consistent subsets of the initial commitments. Enlisting all commitments as a theory always yields an axiomatic base, though it may not be minimal. As only axiomatic bases in the commitments themselves may serve as theories, changes in the commitments are kept at a minimum, and perfect fit is achieved with very low effort.

The streamlining procedure obviously falls short of elaborate accounts of RE, but it approximates the overly simplistic depiction of the state of coherence in RE and the process of working towards it. Still, it would be rather uncharitable to ascribe it to any author that takes part in the debate about RE. Instead, the streamlining procedure will serve us as a *baseline*. In contrast to elaborate accounts of RE, streamlining involves theoretical virtues frugally. There is consistency, and involving minimal axiomatic bases is a very crude move simplicity as axiomatic economy. If there is potential for systematisation within the commitments, it is exploited. However, the selection of more broadly scoped theories, which may have higher revisionary potential, is blocked by streamlining.

Why is (Conservativity) problematic for RE as an account of justification? The value of conservatism as a methodological principle depends on whether there is something worth keeping in the first place. Thus, (Conservativity) gains traction as an objection in the presence of epistemically deficient inputs, such as biases or prejudices that should not be kept. If RE is too conservative, it may not eradicate deep-rooted prejudices or erroneous views (e.g., Singer, 1974, 516; Brandt, 1985, 7), and thus, transfer epistemic deficiencies from inputs to outputs. The idea of preserving deficiencies through RE is called "garbage in – garbage out" (Jones, 2005, 74), and it has been made prominent by Stich and Nisbett (1980), where they discuss the example of the gambler's fallacy. The fallacy can be stated as a principle:

(f) In a fair game of chance, the probability of a given sort of outcome occurring after $n+1$ consecutive instances of non-occurrence is greater than the probability of its occurrence after n consecutive instances of non-occurrence. (Stich and Nisbett, 1980, 192)

Assume that an agent commits to stick betting on the number 23 in a game of roulette, because they formed a belief c that "23 is likely to occur", since it has not occurred for a long time. In this case, f could act as an element in our agent's theory that accounts for their belief c. It is plausible that these propositions can be held as the result of a streamlining procedure. f axiomatises various commitments of the same sort as c about other games. Thus, f accounts for these commitments and consistency obtains as well. The agent may find themselves in a state of equilibrium, weakly characterised as fit between their commitments and a theory consisting of f. However, from the viewpoint of probability theory, f is clearly wrong, and hence, cannot be justified. This would constitute an example for (Weakness): Being in a state of RE is not sufficient for justification. However, this alleged state of RE of

the gambler's fallacy is incomplete and too narrow. We will reconsider the example in a wider context shortly.

Before that, there are additional consequences of a weakly characterised state in terms of fit for RE processes. Consistency consideration for removing inconsistencies may not promote a determinate choice: An inconsistent set of propositions (e.g., $\{p, p \rightarrow q, \neg q\}$) may have multiple maximally consistent subsets ($\{p, p \rightarrow q\}$, $\{p, \neg q\}$, and $\{p \rightarrow q, \neg q\}$), but consistency does not tell us which set to choose, e.g., by ranking the candidates. The same holds for adjustments in view of account. There may be multiple theories that account equally well for the current commitments.

Without further guidance, the adjustment of theories and sets of commitments is underdetermined. Some authors take this to be a flaw of a method of justification, which is supposed to provide the basis for comparative evaluation and rational grounds to prefer a theory over another (e.g., Little, 1984, 384; Haslett, 1987, 307). This brings us in the vicinity of practical weakness problems of RE (e.g., Arras, 2007 or Strong, 2010).

3.3 No-Convergence

For the time being, let us focus on the predominant picture of a process of equilibration leading to a state of equilibrium. Many descriptions of RE make clear that RE does not start from nowhere. Inputs need to be provided to get the process of equilibration off the ground. But if agents set out from different starting points, how do the outputs of their equilibration processes relate? Do they reach the same output, or outputs that are sufficiently similar, such that we may speak of convergence?

The questions surrounding convergence in RE have emboldened critics who object to RE on the basis of a suspected lack of convergence or even the fostering of disagreement. No-convergence objections to RE are a prominent line of criticism that can be traced to the early replies to Rawls (e.g., Singer, 1974, 494), and the objection has been urged again and again since that time.

Most critics proceed by arguing that substantial differences in the starting points survive the process of equilibration and are preserved in the state of equilibrium because RE is overly conservative (Section 3.2) Other critiques do not rely on such initial differences, insisting instead that differences can arise during the equilibration process leading to divergent equilibrium states. In this regard, Bonevac (2004) and McPherson (2015) argue that equilibration

processes are *path-dependent* due to the order of operations or the underdetermination of admissible adjustments.

The spectre of no-convergence is presented as a problem for the justificatory power of RE for various reasons, all of which may be summarised by the worry of Kelly and McGrath (2010) that RE is too *weak* as an account of justification. Divergent outputs reveal that RE has overly pluralistic implications. In its most extreme voicing, RE is suspected to border upon an "anything goes" relativism (de Maagt, 2017, 450; Haslett, 1987, 311). If the justificatory power of RE is staked upon its ability to produce epistemically desirable features that are commonly understood to be at odds with divergence, e.g., moral objectivity (de Maagt, 2017), then this is a serious problem. Moreover, critics fault the method of RE for not being useful in practice. Given the possibility of divergent equilibria, RE supposedly does not offer any means to resolve disagreements (Brandt, 1979, 22; Little, 1984, 383; de Maagt, 2017, 451). Finally, a lack of convergence may put doxastic justification out of reach for RE, and force us to adopt RE as an account of personal justification (see Section 2.3).

Proponents of RE take the threat of no-convergence seriously. Extensions of the simplistic idea of RE, such as the inclusion of background theories intended to lead towards a conception of "wide RE", can be seen as an attempt to make disagreement more "tractable" (Daniels, 1979, 262). Still, Tersman (2018, 7) finds no-convergence to be the "most troubling" objection.

Due to the highly general level at which it has tended to be discussed, RE has largely remained as a metaphor and thus presented an elusive target for objections. Consequently, the treatment of convergence in RE and objections to it remained vague as well. At best, critical stances are based on plausibility considerations that draw from the formal framework of belief revision theory (Bonevac, 2004), or from the Bayesian literature (Kelly and McGrath, 2010). However, these considerations rest on general frameworks of belief change and do not stem from precise, formally worked out accounts of RE.

Here, I take up three aspects of convergence that surface in the debate about RE. We can frame them as questions: Does RE yield a unique output? Does RE promote agreement? Does RE allow for "anything" goes? If the no-convergence objection to RE apply, we should expect negative answers.

3.3.1 Does Reflective Equilibrium Yield a Unique Output?

Convergence to a unique output is the most straightforward entry point to explore convergence in an RE setting. Rawls (1999a, 44) raises the question of unique outputs, but refuses to speculate about it.

In a forceful attempt to show that RE is too weak as an account of justification, Kelly and McGrath (2010, 337) distinguish between *intrapersonal* and *interpersonal* convergence, i.e., whether i) an individual agent with a single starting point, or ii) a group of agents with different starting points reach a unique output, respectively. Uniqueness in the intrapersonal case is a necessary but insufficient condition for interpersonal convergence.

Note that we could also subsume intrapersonal convergence as a special case under interpersonal convergence in a group of agents that share the same starting point. In this case, the question is whether agents can reach different outputs even though they set out from the same starting point.

Kelly and McGrath (2010) grant intrapersonal convergence for the sake of argument, and reject interpersonal convergence subsequently.

3.3.2 Does RE Promote Agreement?

Instead of the uniqueness condition, we might look for a more lenient understanding of convergence in terms of "agreement" and its cognates. These notions are already in use in the literature on RE. Daniels (1979, 274) relates agreement to objectivity and convergence. Nielsen (1982, 293) describes RE as a method to achieve progress from disagreements about some initial commitments to intersubjective agreement. DePaul (2013, 4474) suggests that wide RE offers the means to achieve a "greater degree of agreement" among agents. Taking a critical stance towards RE, multiple outputs fail to converge if they do not fall into "a cluster of similar theories" (McPherson, 2015, 663), or if they are "different" (Singer, 1974, 494), "radically different" (Kelly and McGrath, 2010, 339) or "conflicting"(de Maagt, 2017, 450).

Unfortunately, "(dis)agreement" and its cognates are highly vague notions. On many occasions, they remain undefined, and gradual and categorical readings are not distinguished from one other.[6]

Let us assume for the moment that we have a gradual notion of agreement at hand that is applicable to groups of inputs and outputs.[7] We can compare

[6] A notable exception is (Tersman, 1993).

[7] I will operationalise a gradual notion of agreement in Chapter 11.

initial and output agreement for a group of inputs and their resulting outputs. We may speak of convergence *to some extent* if there is more agreement among the outputs than initial agreement among the inputs.

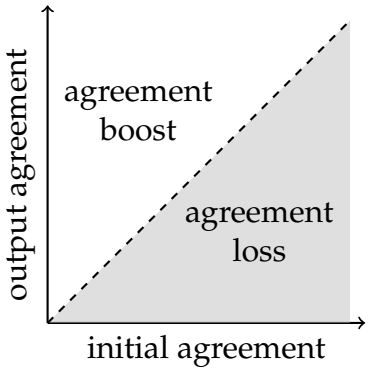


FIGURE 3.2: Agreement among groups of inputs (horizontal axis) and groups of outputs (vertical axis). Agreement increases in the directions of the arrows. The diagonal, dashed line indicates parity between initial and output agreement.

Figure 3.2 displays the basic setting to spell out convergence in terms of initial and output agreement. The dashed line indicating parity between initial and output agreement separates the space into two regions. Convergence to some extent comes about if output agreement is higher than initial agreement in the upper, non-shaded area. In the lower, shaded region, output agreement does not exceed initial agreement or may even be lower than it.

Note that we could also convey convergence to unique outputs as a limiting case in this setting. If full agreement is reached if and only if the inputs converge to a unique output, then convergence to unique outputs would be a horizontal line at the very top of Figure 3.2.

If RE fails to yield convergence in terms of increasing agreement, we should expect to see that RE ends up in the shaded region of the above figure in many cases.

3.3.3 Does Reflective Equilibrium Allow for "Anything Goes"?

Sometimes, RE faces the charge of "anything goes", which takes the no-convergence objection to the extreme. The worry is that the weakness of RE is so pronounced that virtually anything could be justified as (an element of) an RE output. Surveying the literature on RE reveals that we should distinguish at least between two claims about "anything goes" which are directed against RE on two different levels. On the level of sets of sentences, the objection goes that there might be as many outputs as there are inputs (de Maagt,

2017, 450). More precisely, the claim seems to be that the number of different sets of initial commitments is roughly equal to the number of different sets of resulting commitments. Other authors discuss "anything goes" on the level of individual sentences, which amounts to the following claim: For every belief p that is justified to some degree in light of cohering with a set of beliefs, there is another set of beliefs for which the negation of p is equally well justified (Tersman, 1993, 103) (see also (Elgin, 1996, 142)). "Anything goes" on the level of sentences is more fine-grained than on the level of sets because the former could occur even in the absence of the latter. Even if RE evaded "anything goes" on the level of sets by producing only few outputs, the outputs could still allow for "anything goes" on the level of sentences.

3.4 Addressing the Objections

Reactions from RE proponents indicate that they the objections are taken seriously. For example, Rawls (1974, 288) or Scanlon (2003, 150) take up (Conservativity), and Tersman (2018, 7) considers (No-Convergence) to be "most troubling". Have the objections to RE been addressed by proponents of RE with convincing rejoinders? The fact that the earliest objections found in the "first wave" (Hare, 1973; Singer, 1974, Lyons, 1975), have been restated or elaborated even after updated accounts of RE (e.g. Daniels, 1979) in more recent works (e.g., Brandt, 1985; Arras, 2007; Strong, 2010, Kelly and McGrath, 2010; McPherson, 2015; McGrath, 2019; Dutilh Novaes, 2020) suggests otherwise. In this section, I aim to present two important elaborations of RE in order to escape (Weakness) and its subordinate objections: including background theories, or imposing constraints on inputs. These amendments did not go unnoticed by critics, and new as well as old issues have been raised. I do not aim for a conclusive assessment of whether proposed additions save RE from falling prey to (Weakness), or whether the updated objections are defeating, which would require a much more detailed treatment. For the purpose of this project, we can take the stalemate in this debate as a motivation to explore the prospect of strengthening RE with theoretical virtues as an additional approach to defend RE.

Background Theories Background theories are a well-known extension of RE made prominent by Daniels (1979), resulting in what is called "wide" RE. Background theories, e.g., theories about persons in Rawls' work, enter RE consideration as a third element beside commitments and theories

(Daniels, 1979, 258). RE states are characterised as coherence among a triple of commitments, a theory and relevant background theories, and RE process may additionally involve adjustments in background theories.[8] According to Daniels, arguments inferred from background theories "bring out the relative strengths and weaknesses of the alternative sets of principles", i.e., they provide the basis for evaluating and selecting competing theories in RE processes independent of their fit with commitments (Daniels, 1979, 259).

How does the addition of background theories to RE help addressing the objections? It is the function of a background theory to prevent that the principles in (foreground) theories are not "mere accidental generalizations" of commitments (Daniels, 1979, 259). Moreover, background theories counteract the overly conservative streamlining of starting points as they are supposed to have broader scope than the commitments (Daniels, 1979, 259). Consequently, an agent would be required to extend her commitments in order to increase coherence, and by this, learn something new. Background theories allow for additional and more drastic revisions of commitments than narrow RE (Daniels, 1979, 266). All of this can be seen to counteract (Conservativity).

Concerning (No-Convergence), Daniels puts forward the idea (supported by two examples) that disagreements about background theories are more "tractable" and "manageable" than disagreements about commitments or elements of theories (Daniels, 1979, 262–263). If agents agree on their background theories, their adjustments in an RE process are guided by the same considerations, which makes converging outputs more plausible. Daniels questions whether background theories warrant convergence. In face of disagreement about commitments in support of background theories (and hence, probably also about background theories themselves), better tractability by background theories is not guaranteed (Daniels, 1979, 264).

Moreover, Daniels expects background theories to be less prone to epistemic deficiencies."But it may also be that the agreement is found because some of the background theories are, roughly speaking, true — at least with regard to certain important features" (Daniels, 1979, 272). Thus the involvement of background theories may help to eradicate epistemically deficient inputs during an RE process, which is guided by the background. Moreover, RE states are less likely to be epistemically deficient since they are required to be coherent with background theories, too.

[8] He rejects the idea that specific elements of RE are immune to revision (Daniels, 1979, 264), which is another source of (Conservativity).

Let us have another look at the illustrative example of the alleged state of equilibrium between the gambler's fallacy and commitments to bet on numbers that have not occurred for a long time in a game of chance. The crucial point, which is missing in the example of Stich and Nisbett (1980), is that the agent, at this point, is in a very *narrow* state of RE. What if the agent would consider the addition of basic probability theory to his background? Clearly, the agent would no longer be in a state of RE due to incoherence. f is inconsistent with insights from probability theory, in particular statistical independence of events such as coin tosses or spins of a roulette wheel. In face of this inconsistency, the agent has various options to resolve the conflict. On the one hand, they could remove or revise f (e.g., by negating it), followed by giving up or revising the belief they are committed to. On the other hand, the methodology of wide RE does not prevent an agent to revise his background theory about probability to re-establish coherence. The problem with this move is that it would amount to the monumentally task of reforming probability theory. However, probability theory itself is presumably part of a quite stable equilibrium. Not only would a revision of probability theory have to be coherent, it would also need to be successful in a wide range of other cases. Thus, by attempting to revise their background theory, the agent is at high risk of performing an ad-hoc manoeuvre in order to save their commitment and principle.

The addition of background theories to RE has been taken into account by opponents of RE and lead to updated criticism. One target of criticism is Daniels' *independence constraint*, according to which background theories should be supported by a set of commitments, which is disjoint from the commitments systematised in the (foreground) theory (1979, 260).

(Little, 1984) and Haslett (1987) reject partitioning the commitments into those for narrow RE and those to support the background theories. Even narrow RE should take *all* commitments into consideration. Strong (2010) argues that independent commitments in support of background theories can also be subject to historical accident or bias, and hence wide RE may still result in a mere systematisation of biases failing to correct errors. Thus, the issues of conservativity and deficiency re-enter at the level of background theories: An epistemically deficient background (e.g., due to bias) is of no help to counteract the streamlining of prejudices in the foreground. So, the worry is that the inclusion of background theories merely shifts the problems to another level.

Moreover, opponents of RE do not share Daniel's expectation that there is a tendency towards more agreement about background theories (e.g., Strong,

2010, 131). In face of widespread philosophical disagreement about all kinds of theories, it is unclear whether disagreement about commitments and elements of theories are made more tractable or manageable by the inclusion of background theories. According to this view, (No-Convergence) cannot be ruled out by relying on background theories.

I also observe an expansionist strategy among proponents of RE (e.g., Tersman, 2018, 7; Scanlon, 2003, 152f) to take known disagreement among different agents into account as it may disrupt less wide equilibria. However, Strong (2010, 130) re-raises the point of deeply ingrained bias since the consideration of many viewpoints does not eradicate it, and de Maagt (2017, 458) criticises that this move makes RE vacuous and impractical as it reduces RE to reasoning about a subject matter in general.

Constraints on Inputs The role of the epistemic standing of inputs, especially initial commitments, is probably one of the most controversially discussed points about RE. The above exposition of objections revealed the involvement of epistemically deficient inputs at various stages: In combination with (Conservativity), epistemically deficient inputs support "garbage in - garbage out", and they can also be seen as a source of disagreement, which ultimately leads to (No-Convergence).

The most prominent idea in order to "sanitise" inputs stems from Rawls, namely that only *considered* judgements should enter the process of RE. He characterises considered judgements as follows:

> [...] they enter as those judgments in which our moral capacities are most likely to be displayed without distortion. [...] For example, we can discard those judgments made with hesitation, or in which we have little confidence. Similarly, those given when we are upset or frightened, or when we stand to gain one way or the other can be left aside. All these judgments are likely to be erroneous or to be influenced by an excessive attention to our own interests. (Rawls, 1999a, 42)

Allowing only considered initial commitments to enter an RE may reduce the risk of running into the issues raised by the objections: (Conservativity) is problematic if it is coupled with epistemically deficient inputs, which are filtered out by consideration before they can enter a process of equilibration, or even be preserved in a state of equilibrium. Concerning (No-Convergence), the filtering of initial commitments by consideration may lead to more agreement among inputs, and also to less disagreement among outputs. Rawls

(1951, 182f) includes stable agreement among competent judges as a characteristic of a considered judgement. This characteristic is absent from his classical exposition of RE (Rawls, 1999a, 18f, 42–44), and mentioned explicitly only much later (Rawls, 1999a, 508).

So the crucial question is the following: Is the filter of consideredness sufficient to prevent epistemically deficient initial commitments? Critics answer in the negative and McPherson (2015) formulates the discontent cogently while citing (Kelly and McGrath, 2010, 346–348):

> The dispositional criterion of considered moral judgments means that the method can endorse intuitively monstrous judgments as appropriate starting points for normative theorizing, provided these judgments are held with the right sort of dispositions.
> (McPherson, 2015, 663)[9]

Thus the dispositional characterisation of considered judgements fostered the dissemination of "counterexamples", i.e. invitations to imagine specific situations, in which agents hold epistemically deficient commitments despite having the correct dispositions.

One reaction in defence of RE would be to demand a firmer epistemic standing of inputs, for example given by a normative characterisation of considered judgements. Opponents claim that this move is problematic since it threatens to make RE uninformative or uninteresting as a method of justification. Kelly and McGrath (2010, 353–354), for example, argue that if the method of RE requires already justified or highly credible beliefs, the method becomes uninteresting, because it shifts the focus to what makes starting points reasonable and how we grasp such facts.

Somewhat related to this point is the criticism that strengthening the input moves RE from a purely coherentist account of justification to moderate foundationalism. However, this line of criticism ignores the third option, namely to base RE on a weakly foundationalist epistemology (see Section 2.3).

Outlook: Theoretical Virtues Background theories and constraints on inputs indicate that RE possesses the means to address objections, but above overview suggests that it may leave critics wanting. Background theories

[9] He applies a similar point to the operations of adjustments in RE processes, which are also dispositionally characterised by Rawls (McPherson, 2015, 663–664).

and constraints on inputs would deserve closer attention to asses their prospects and limitations, but I suppose that we have enough motivation to turn to an under-explored line of defence for RE.

The idea of considering theoretical virtues in RE allows to put these long-standing discussions about other aspects of RE aside (at least provisionally), if the objections can be addressed from an alternative angle. If the involvement of theoretical virtues helps to deal with (Conservativity) and (No-Convergence), they undermine two line of thoughts that lead into (Weakness). In this sense, theoretically virtues would contribute more or less indirectly to a strengthening of RE as an account of justification.[10]

Concerning (Conservativity), looking back at the depiction of elaborate RE in Figure 2.1, reveals that the discussion of objections and the replies so far does not involve the right-hand side of doing justice to theoretical virtues which are configured in light of pragmatic-epistemic objectives. Theoretical virtues exert pressure to come up with a virtuous theory. This "progressive" force of doing justice to theoretical virtues is transferred by the attractive force of fit from theory to the commitments. Hence, doing justice to theoretical virtues pulls in the opposite direction of the "conservative" pull of respecting input commitments. One way to think of this tensions is systematisation. Initial commitments represent an agent's unsystematic collection of views about a subject matter at the outset of RE. In turn, theoretical virtues spell out systematicity. Whether this thug of war between forces in RE goes in one direction rather than the other depends on relative weighing of respecting the input, fit and doing justice to theoretical virtues. RE state needs to strike a balance.

Moreover, doing justice to theoretical virtues goes against an RE process being a mere streamlining procedure (as described in Section 3.2). Streamlining includes some virtues, e.g., consistency, but it lacks other virtues that are generally relevant to systematisation, e.g., broad scope. We can expect revisions during an process of equilibration process that go beyond establishing mere fit. Thus, theoretical virtues offer an approach to deal with (Conservativity). Doing justice to theoretical virtues provides incentive to revise commitments (mediated by fit) more substantially than streamlining.

[10]One might also argue for a more direct strengthening of RE with theoretical virtues. If an RE state is supposed to include a virtuous theory, and theories are justified qua exhibiting virtues, the justificatory power of RE stems from them. This direct strengthening of RE as an account of justification would turn on the assumption that theoretical virtues are intrinsically valuable from the viewpoint of justification.

Concerning (No-Convergence), theoretical virtues provide more structure to theory adjustments in an equilibration process. They may filter candidate theories according as necessary requirements (e.g., consistency) or order them according to their virtuousness (e.g., simplicity). This restricts the available options reducing the number of paths an agent could take from their starting point. Moreover, the evaluation of whether a state qualifies as reflective equilibrium also involves the assessment of its theory's virtuousness in view of available alternatives (condition 4 on page 22).

Involving theoretical virtues does not guarantee unique outputs, as there may be multiple equally virtuous theories. But the restriction of admissible theory candidates may still lead to some convergence in terms of increased agreement among outputs.

Whether the involvement of theoretical virtues are successful in addressing (No-Convergence) also turns on the assumption that agents agree on the configuration of theoretical virtues. Otherwise, differently configured theoretical virtues might constitute another source of divergence. For example, if agent_1 prefers simple theories over broadly scoped ones, and agent_2 prefers scope over simplicity, they may reach quite different states even if their starting points are similar in other aspects. However, the standard answer in support of RE applies here as well: Reaching different equilibria due to different configurations is not a problem per se. Applying elaborate RE forces agents to be explicit about their details of their epistemic situation and disagreements that can be tracked to differences in epistemic situations may disrupt provisional equilibria towards even wider reflective equilibrium. Whether there is agreement about configurations of theoretical virtues is a question that is addressed empirically in (Schindler, 2022) for different groups of scientists with results that allow to be moderately optimistic.

Appendix

A.1 A Map of Some Objections

Figure A.1 is a schematic illustration of how objections to RE relate to each other.

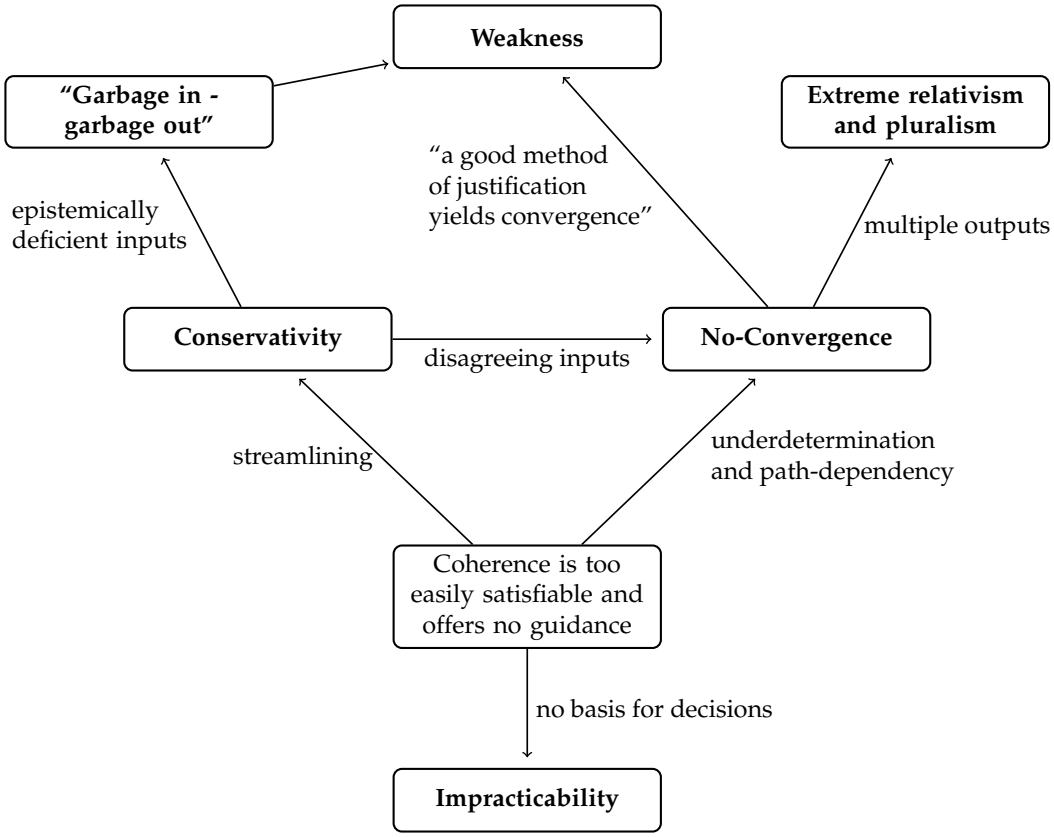


FIGURE A.1: Some objections discussed in Chapter 3 and their interrelationships. The arrows indicate the directions in which different line of thoughts lend support to other objections.

Chapter 4

Theoretical Virtues

Many accounts of RE distinguish between commitments and theories, and involve the idea that the latter should systematise the former. But what exactly is systematisation? Moreover, equilibration processes or equilibrium states involve the evaluation of theories. Which theory adjustment is appropriate given the current commitments of an agent? What are the repercussions on theories if equilibrium states need to be as good as any available alternative? A plausible methodological advice voiced in elaborate accounts of RE is to include the considerations of theoretical virtues. This is a welcome addition to RE, as the inclusion of theoretical virtues in RE may constitute untapped potential to address the objections against RE. Let us move away from the highly general level of methodological advice, and spell out the role of theoretical virtues in RE in more detail. A promising starting point lies in philosophy of science, where the appraisal of scientific theories plays a prominent role.

The aim of this chapter is to develop an understanding of theoretical virtues from philosophy of science that allows to integrate them into RE. Ultimately, I propose how to render theoretical virtues fruitful for RE whilst moving from the methodological advice to an applicable method.

I organise the work towards this goal as follows. In Section 4.1, I survey theoretical virtues in philosophy of science. This helps to get a grasp of the many roles that theoretical virtues play in philosophy of science, and fix some terminology. There are some attempts of systematising theoretical virtues from which we can draw some helpful distinctions. However, longstanding issues remain unsettled, and there are no ready-made solutions that we could import to RE.

Next, I turn to theoretical virtues in the literature about RE in Section 4.2. I track theoretical virtues to classical and elaborate accounts of RE, which reveals an abundance of elements that are similar to those in philosophy of

science. Moreover, I present and discuss a critical stance towards to prospects of including theoretical virtues in RE.

Against this background, I formulate a proposal in Section 4.3. In order to integrate theoretical virtues fruitfully into a cognitive enterprise they need to be configured in view of pragmatic-epistemic objectives, that guide the selection, specification, weighting and aggregation of virtues. I suggest to focus on virtues which are related to coherence as a universal but vague objective of RE. By going beyond mere consistency and fitting together, theoretical virtues can render the notion of coherence more substantive.

4.1 Theoretical Virtues in Science

Which scientific theory should you accept when data alone does not dictate a choice? What makes an explanation the best among others? How should we demarcate science from other cognitive enterprises? Attempts to answer such questions comprise decades of research in philosophy of science. At first glance, there is some minimal common ground: Many authors deploy epistemically desirable features of theoretical elements – theoretical virtues– in developing their points. A closer look, however, reveals a dire situation: There is terminological disparity, endless lists of more or less vague virtues, and longstanding issues remain to be addressed. Sparse attempts of systematisation are implicitly shaped by specific philosophical views, and there are very few overview articles on this subject matter. Hence, I aim to provide a broad overview before moving to RE. This bears the potential to provide interesting connecting points for further topics that arise in philosophy of science as well as in RE, such as objectivity or understanding.

The Many Roles of Theoretical Virtues Theoretical virtues permeate a wide range of philosophical discussions about science presented as a tour d'horizon in the following paragraphs. For an extensive bibliography of theoretical virtues in philosophy of science, see (Schindler, 2020).

Famously, Kuhn (1977) discusses theoretical virtues for theory choice in relation to objectivity, a topic that also appears in (Hempel, 1983, 1988) and many others up to this day (e.g., Heron, 2020). In the vicinity of objectivity, we might stumble upon the ideal of value-free science, claiming that specific stages of scientific inquiry should be kept free from the influence of ethical, social or political values (sometimes summarised as "contextual" values.

(Lacey, 1999) is an important contributions arguing in favour of the ideal, but it has also come under attack (e.g., Longino, 1996, or Douglas, 2009).

Next, theoretical virtues figure in the demarcation of science from mathematics and logic, as well as from metaphysics. Popper (1959b) proposes falsifiability as central criterion.[1] Hoyningen-Huene (2013) demarcates scientific knowledge from other forms of knowledge, especially everyday knowledge, arguing that the former is more systematic than the latter.

For explication, i.e., the transformation or replacement of a vague or prescientific concept with an exact one, Carnap (1950, 7) includes virtues that serve as requirements of adequacy. Lipton (2004) provides an account of inference to the best explanation. So-called explanatory virtues make an explanation "lovelier", i.e., the explanation provides more potential understanding (Lipton, 2004, 58–59, 122–123). For an account that ties theoretical virtues to understanding of theories by scientists, see (De Regt, 2017, Chapter 2).

The probably most fiercely debated aspect of theoretical virtues in philosophy of science concerns the question, which theoretical virtues are related or conducive to truth (scientific realism) or empirical adequacy (antirealism). Both realists and antirealists include theoretical virtues in their accounts, but differ starkly with respect to the role of virtues. For theoretical virtues in an influential antirealist account, see (van Fraassen, 1980). On the side of realism, Psillos (1999, 165–169) involves theoretical virtues, and Schindler (2018) provides a recent defence that revolves entirely around theoretical virtues.

Occasionally, the underdetermination of scientific theories by evidence is brought forward against realism. Underdetermination holds, if there are multiple, empirically equivalent, but incompatible, theories for every body of evidence, and empirical adequacy is the only epistemically admissible virtue for choice. However, the prospects of attacking realism by underdetermination or escaping it by means of including additional theoretical virtues has been put into doubt (Kukla, 1994; Tulodziecki, 2012). According to another reading, underdetermination is the idea that we cannot test hypotheses in isolation from auxiliary hypotheses or background beliefs. Consequently, one need not to reject the hypothesis in face of failed tests, but revise the background. What could provide further guidance in this situation? Duhem (1954, 216–218) proposes to use "good sense", i.e., to deliberate whether we

[1] A theory is falsifiable if and only if it unambiguously divides the class of basic statements, i.e., observational reports, into a subclass of potential falsifiers, with which it is inconsistent, and a subclass of statements, which are permitted since they do not contradict the theory. (Popper, 1959b, 65–66). Thus, falsifiability presupposes consistency, because an inconsistent theory would allow for any partitioning of statements.

should discard principles of a "vast and harmoniously constructed theory", go for slight modifications of details, adhere to repairs that complicate the theory or allow for corrections "to construct simple, elegant, and solid system"(Duhem, 1954, 217). Quine (1955) and Quine and Ullian (1978, 66–80) point out virtues that speak in favour of a hypothesis' plausibility.

Last but not least, Lewis (1973, 73–77) includes two virtues of deductive systems in his account of natural laws, namely simplicity of its axiomatisation and strength or information content. He also remarks that these virtues can conflict and thus may be in need of balancing.

I terminate my survey at this point, even though there would be much more material to cover. For instance, there are applications of theoretical virtues to more specific problems, e.g., curve fitting (Forster and Sober, 1994) or justification of mathematical axioms (Heron, 2020). Only recently, theoretical virtues became the object of more empirical research in philosophy and science. Schindler (2022) presents a quantitative study that compares views of scientists from natural and social sciences, as well as philosophy. Among other insightful results concerning the epistemic status of specific virtues, an important observation is widespread agreement among all three groups about how to order five commonly stated virtues. Mizrahi (2022) finds mixed results in a large corpus analysis by means of text mining. The relative frequency of published scientific texts that explicitly mention virtues is rather low.

We can draw at least two lessons from the observation that theoretical virtues permeate philosophy of science. First, there is shared appreciation for theoretical virtues among philosophers of science. This is enough motivation to attempt a transfer. If theoretical virtues are widely appreciated and versatilely involved in philosophy of science, why should they not prove to be beneficial for RE as well? After all, there is a structural parallel between science and RE, even if it might turn out that the analogy is very superficial.[2] Evidence or commitments, respectively, are contrasted with theories. In both cases, theory choice or change stand in need of guidance for the evaluation of candidate theories.

Second, the diversity of pursued goals and diverging points of view sheds some light on the question why there has been very little progress towards

[2]For a warning not to overstretch the analogy between scientific and RE methodology, see Sayre-McCord, 1996, 142, or Cummins, 1998. However, this pertains to treating commitments as some kind of evidence, against which theories are to be tested. However, the analogy between the roles of theoretical virtues in science and in RE is not affected by this, and in my view, not problematic.

a unified account of theoretical virtues on the highly general level, on which above contributions to philosophy of science operate on. I delve into the second point in the following sections. First, I fix terminology that I want to use to navigate diverse vocabularies providing some clarificatory remarks. Second, I have a look at some attempts to systematise theoretical virtues, and I see what we can learn from them. Finally, I present longstanding issues of theoretical virtues.

Terminology In view of the wide range of debates in philosophy of science that involve epistemically desirable features of theories, it may not come as a surprise that terminology diverges substantially. I personally like to speak of "theoretical virtues" but this is by far not the only option. Other popular terms in place of "virtue" are "criteria" or "values" (Kuhn, 1977), "desiderata" (e.g., Hempel, 1988), "requirements" (Popper, 1959b, 72–73; Carnap, 1950, 5–8) "standards" or "virtues" (Quine and Ullian, 1978). Note that a terminological decision to use one notion rather than another may be accompanied by many implicit subtleties. A clear cut criteria, a desideratum allowing for degrees, or a value held by an agent express substantially different roles that a theoretical virtue may play.

Theoretical virtues are commonly ascribed to "scientific representations" comprising theories, hypotheses, or models (Hirsch Hadorn, 2018, 321).[3] Popper (1959b), Lewis (1973, 87) or Sober (2002, 13) discuss the virtues of hypotheses, and for models, see (Forster and Sober, 1994, 14–15). In addition, virtues figure in evaluating explanations (Lipton, 2004) or even beliefs (Lacey, 1999, 45). Whether theoretical virtues are ascribed to theories, hypotheses or models, may influence the selection of relevant virtues. Hypotheses, for example, are supposed to have *testable* consequences, a model *predictive accuracy* with respect to one aspect of a target system, and a theory may aim at the *unification* of yet diverse phenomena. For empirical results that stand in support of this point, see (Mizrahi, 2022, 18).

I use "theoretical virtue" for epistemically desirable features of theories to circumvent confusion with *intellectual* virtues of agents (e.g., courage, humility or epistemic justice) that are discussed in the field of virtue epistemology (for an overview, see Turri, Alfano, and Greco, 2021). This also avoids other

[3]Often, authors rely on an informal understanding of such representations. For example, it is not discussed whether the structure of scientific theories is to be understood syntactically (as a set of theoretical sentences), semantically (as a class of models), or pragmatically (as a complex of formal and non-formal components). For an overview of these views, see (Winther, 2021).

commonly used terms in this place, such as "epistemic" or "cognitve" that can be understood very narrowly or very broadly. I am reluctant to use these terms, since they also foreshadow a partition of virtues in those, which qualify as "epistemic" or "cognitive" and those, which do not.

Another clarification concerns the use of *categorical* and the *comparative* readings of theoretical virtues in evaluating theories. On the one hand, theories can "have" or "exhibit" theoretical virtues, which is a matter of yes or no. Consistency, for example, is often treated in this way. If we assume that we have no means of measuring the number or severity of contradictions, consistency rests upon the absence of contradictions. On the other hand, theories are frequently compared to each other with respect to theoretical virtues. A theory may, for example, be simpler, have less explanatory power, or be equally fruitful than another theory. In this case, theories instantiate virtues in varying degrees that allow for ordering relations between them: " ... performs at least as well as ... with respect to virtue v ". At this point it is important to note that every comparatively read theoretical virtue gives rise to an ordering relation on its own, which leaves open the question whether and how we can arrive at an aggregated relation of overall betterness " ... is overall more virtuous than ... ".[4]

We can transform a comparative reading of a theoretical virtue into a categorical one by setting a threshold and ascribing the virtue to every theory that passes the threshold by instantiating the virtue to a sufficient degree.[5]

Systematising Theoretical Virtues The proliferation of theoretical virtues results in long and unordered lists. Divergent use of terminology worsens this situation. As a rule of thumb, many authors tend to state approximately five elements explicitly on their non-exhaustive lists of virtues (e.g., Kuhn, 1977; Quine, 1955; Quine and Ullian, 1978; Schindler, 2018). An outlier to this trend is Lacey (1999), who collects dozens of elements on his list.

Is it possible to impose more structure on theoretical virtues? Are there classifications that reveal shared features across different proposals or insightful interrelationships? While there are some helpful distinctions, other aspects remain controversial to the point of becoming a hindrance. Most notoriously, the question of which theoretical virtues are truth-conducive is fiercely disputed.

[4] I will elaborate on this topic in Chapter 6.

[5] There are some ideas on what could count as non-arbitrary standards to determine sufficient degree of manifestation of a cognitive value (Lacey, 1999, 61–65).

Laudan (1984, 2004) introduces the label of "cognitive" values comprising the values that are "constitutive of science in the sense that we cannot conceive of a functioning science without them"(Laudan, 2004, 19). Within the category of cognitive values, Laudan distinguishes "epistemic" (related to truth or probability) from "non-epistemic values", which are called "pragmatic" by other authors (e.g., van Fraassen, 1980, 88).[6]

On Laudan's basis, Douglas (2009, 2013) attempts to provide philosophical justification to values and resolve alleged tensions among them with a systematisation. She proposes two distinctions (Douglas, 2013, 798–799). First, there is a difference between *minimal criteria*, i.e., features that a scientific theory has to instantiate, and *ideal desiderata*, i.e., traits of a theory that are valued by scientists even if they are not fully instantiated. Second, a value may pertain to a theory *on its own* or *in relation to evidence*. Since both distinction are individually applicable to a value, Douglas can carve up four groups:

	theory on its own	**in relation to evidence**
minimal criteria	internal consistency	empirical adequacy
ideal desiderata	scope	unification
	simplicity	novel prediction
	potential explanatory power	precision

TABLE 4.1: Douglas (2013) groups theoretical virtues according to whether they pertain to a theory on its own or to the relation of a theory to evidence, and whether they serve as criteria or as desiderata.

Douglas supposes that the values from the first two groups (top row in Table 4.1) "are genuinely truth assuring, in the minimal sense that their absence indicates a clear epistemic problem" (Douglas, 2013, 799). The members of the third group (scope, simplicity and explanatory power) are considered to be "strategic or pragmatic values" in as much as they are an "aid to thinking" and facilitate the ease of use. They "give no assurance as to whether the claims that instantiate them are true but give us assurance that we are more likely to hone in on the truth with the presence of these values than in their absence"(Douglas, 2013, 800). She subsumes the values of the third group

[6]The literature provides us with much more such dichotomies, e.g., constitutive vs. contextual (Longino, 1996) or intrinsic vs. extrinsic (Steel, 2010), which I cannot give due consideration.

under the rubric of "fruitfulness". The fourth group (unification, novel prediction, precision) consists of values that again have genuine epistemic import since they provide "assurance against ad hocery" (Douglas, 2013, 801).

How does Douglas' systematisation resolve the tension among theoretical virtues? The minimal criteria from the first and the second group are prioritised, they must be met as necessary requirements of adequacy. A scientific theory is not adequate as long as it does not fulfil these criteria, even though scientist may provisionally work with inconsistent or empirically incompetent theories. Thus, by construction, the required minimal criteria cannot be in conflict with the ideal desiderata, which are optional amenities of theories. Furthermore, consistency is construed as a prerequisite condition of empirical adequacy (Douglas, 2013, 801). Hence, there is no tension between the first and the second group.

The third and the fourth group serve different purposes by dissolving possible conflicts (Douglas, 2013, 804). The former concerns the fertility of theories and their ease of use and the latter provides epistemic assurance to the question of what is the best supported theory or the most reliable knowledge at this point in time. Within the third group, differences may arise, but they have no epistemic import due to their pragmatic nature. In addition, diversity in this group may even benefit science (Douglas, 2013, 802). Finally, some tensions remain within the fourth group.

Douglas divides theoretical virtues into those which are have genuine epistemic import and those which are pragmatic. I think that her distinctions for systematising theoretical virtues are very helpful independent of a epistemic-pragmatic partitioning. After some amendments, I will apply the distinctions to the project at hand, even though I beg to differ with respect to the classification of specific virtues. This may be a consequence of having different goals in mind that motivate the inclusion of theoretical virtues. For example, Bhakthavatsalam and Cartwright (2017) point out that making consistency and empirical adequacy minimal requirements presupposes that science aims at truth. For other purposes of science (e.g., understanding, managing the world) it may be unreasonable to make empirical adequacy a minimal requirement (Bhakthavatsalam and Cartwright, 2017, 449–450).

A shortcoming of Douglas systematisation is the inability to capture virtues concerning the relation of a theory to other theories, for example external consistency (e.g., Kuhn, 1977). To be fair, Douglas is fully aware of those theoretical virtues and she includes them explicitly under unification (Douglas,

2013, 801). She construes them as pertaining to a theory in relation to evidence that supports other theories, so that external consistency is not out of the frame. However, we still lack the ability to evaluate how a theory relates to other theories independent of evidence. For example, take the gambler's fallacy as an element of a vicious theory (see Section 3.2). In this case, we can base our negative evaluation on the external inconsistency inconsistent with axioms of probability theory. This evaluation rests on completely theoretical considerations and does need to involve evidence, which also could hint at something being wrong, e.g., if we observe unaltered chances after long streaks of the same outcome.

Hence, I propose to complement Douglas' systematisation as follows. We keep the distinction between minimal criteria and ideal desiderata, but distinguish between *purely* theoretical virtues, i.e., features of theories on their own (e.g., internal consistency), and *hybrid* virtues of theories in relation to other components of inquiry. Hybrid virtues then can be subdivided in those that pertain to the theory in relation to the evidence (e.g., accuracy), and in relation to other theories (e.g., external consistency).

Note that these distinction also apply to an RE setting that includes a distinction between commitments, theories and background. In order to have a rich set of virtues at our disposal to evaluate theories, we can read hybrid virtues as virtues of theories relative to given commitments. Note that this not a move out of an embarrassment of scarcity. The present survey of theoretical virtues in philosophy of science reveals that it is quite common to count hybrid virtues that relate evidence and theories as theoretical virtues. Thus, I will construe theoretical virtues for RE broadly, including both pure and hybrid virtues.

After this quick detour, let us head back to philosophy of science. Let me illustrate the potential to arrive at a significantly, different systematisations with a proposal by Keas (2018). Building on the work of McMullin (2008) on theoretical virtues, he describes twelve theoretical virtues in detail and proposes a systematisation into four groups. The virtues within a group "sequentially follow a repeating pattern of progressive disclosure and expansion" (Keas, 2018, 2762), displayed in Table 4.2.

He ranks the groups according to their epistemic weight, where he understands "epistemic" in a broader sense including more than the aim of truth, e.g., understanding (Keas, 2018, 2763). Evidential and coherential virtues receive high epistemic weight, the aesthetic virtues are granted modest epistemic value. He includes also includes diachronic virtues that may be used

evidential	**coherential**	**aesthetic**	**diachronic**
evidential accuracy	internal consistency	beauty	durability
causal adequacy	internal coherence	simplicity	fruitfulness
explanatory depth	universal coherence	unification	applicability

TABLE 4.2: Keas (2018, 2762) proposes to group theoretical virtues into four groups (columns) in order to systematise them. Rows from top to bottom follow a sequence of "progressive disclosure and expansion".

to evaluate a theory on a temporal dimension after its initial formation. Diachronic virtues rest upon the virtues from the other groups, and if they are instantiated in a "mature" theory, they contribute to the epistemic value (Keas, 2018, 2788).

A comparison with the proposal of Douglas (2013) reveals striking differences. Douglas, for example, subsumes scope, simplicity, and explanatory power as purely pragmatic virtues under fertility, which is treated by Keas as diachronic virtue with a temporal dimension. In turn, Keas develops an interrelationship of simplicity and unification (introduced by Mackonis, 2013) as aesthetic virtues, whereas Douglas puts them in different groups. I do not think that these differences would vanish if we introduced common terminology and indeed, Keas (2018) argues at some length for the superiority of his systematisation in view of Douglas' systematisation.

Issues of Theoretical Virtues The issues of theoretical virtues, which have been raised by Kuhn (1977), are probably as well-entrenched as his list of theoretical virtues itself. The use of virtues in theory choice faces at least two problems (Kuhn, 1977, 357). First, the virtues, taken on their own, are ambiguous. That is to say, that scientist may legitimately differ about the interpretation of a virtue or its application to a particular case. Take for example simplicity, which has a wide range of proposed interpretations, such as the number of theoretical posits, low mathematical complexity in terms of free parameters, or as an aesthetic feature concerning a theory's elegance. Another source for ambiguity lies in the interdependence of virtues (Kuhn, 1977, 364). Kuhn himself does not offer an example of such dependencies, but the literature provides some ideas, e.g., a reciprocal relationship between simplicity and unification (Keas, 2018; Mackonis, 2013). Second, virtues may conflict when they are applied together, because they may pull in different

directions. If the virtues need to be traded off against each other, scientist may differ in their assignment of relative weights to them.

Consequently, persons committed to the same list of criteria for choice may nevertheless reach different conclusions (Kuhn, 1977, 358). An algorithmic decision procedure, which would conclusively determine choice, is a "not quite attainable ideal" (Kuhn, 1977, 359), because the criteria would need to be stated unambiguously and require an appropriate weight function for their joint application. But according to Kuhn, little or no progress has been achieved for either problem. As such they provide an insufficient basis for a *shared* algorithm of theory choice (Kuhn, 1977, 362).

The attempts to systematise theoretical virtues (Douglas, 2013; Keas, 2018; Mackonis, 2013; McMullin, 2008) certainly achieve some progress for the first issue. Classifying virtues into distinctive groups helps to disentangle otherwise long concatenations of unrelated and vaguely characterised virtues. Concerning the second issue of trade-offs, there is the idea to separate minimal criteria from ideal desiderata (Douglas, 2013). This effectively prevents some trade-offs by granting minimal criteria lexicographic priority over desiderata. The same holds for theoretical virtues that are considered to be prerequisites for others, e.g., consistency for empirical accuracy. As it stands, current systematisations of theoretical virtues do not offer solutions to handle genuine trade-offs.

An additional, interesting point made by Kuhn, which is frequently absent in his reception, concerns a positive feedback loop. Not only do virtues influence theory choice, but theory change may also affect the application of relative weights to virtues (Kuhn, 1977, 365). His example is the loss of qualitative accuracy as a value in the aftermath of accepting Lavoisier's theory of chemistry. Kuhn notes that this feedback loop "does not make the decision process circular in any damaging sense", due to the relative stability provided by the often unconscious and delayed changes of value of small magnitude (Kuhn, 1977, 365). Interestingly, this idea is paralleled in the literature on RE, especially if the authors stress the importance of theoretical virtues in RE. Rechnitzer (2022), Baumberger and Brun (2021), and Elgin (2017), for example, see room to update the weighting (and other aspects of a configuration) of theoretical virtues dynamically during the process of equilibration.

4.2 Theoretical Virtues in Reflective Equilibrium

Many elaborate as well as classic accounts of RE involve theoretical virtues more or less explicitly, which provide us with a considerable amount of virtues that exhibit a notable overlap with virtues in philosophy of science. We could go through the entire literature on RE and meticulously compile a long list of virtues that are mentioned explicitly, or that could be ascribed to authors with some interpretative effort. However, this would not be very useful to spell out precisely the role of virtues in RE. Theoretical virtues involved in RE remain mostly undefined and vague at best, which is in line with the vagueness of idea's that surround RE at its general level of discussion. Listings would have to ignore fine-grained distinctions, contexts, objects of instantiation and functions of virtues as intended by the authors. In addition, a list would not help to identify possible reductive relations among its elements, leaving us with embarrassingly many options, and rendering the upcoming task of integrating theoretical virtues fruitfully into RE even more cumbersome.

Consequently, I will focus on examples in the work of Rawls and Daniels, the more recent and elaborate account of Elgin, and the critical stance of Kappel. Tracking theoretical virtues in the "classics" helps to establish that we should think of theoretical virtues as an integral part of RE, and not just as a new patch for some issues of RE. From the work of of Elgin we can compile the most extensive list of virtues as well as a treatment of their functions for RE. Finally, Kappel's list of epistemic desiderata proves to be a very useful compilation of virtues that are relevant to RE, even though his criticism of RE can be evaded.

I will not approach theoretical virtues from a more general side of philosophy, for example with respect to moral theories (Timmons, 2012).

The Classics: Rawls and Daniels Theoretical virtues surface in Rawls' formative contributions to reflective equilibrium although they never take centre stage. His most explicit treatment of virtues can be found in his early *Outline of a Decision Procedure for Ethics* (1951), which presents some, but not all, components of RE. He proposes an "explication" of considered judgements with a set of principles as a heuristic device to yield reasonable and justifiable principles (Rawls, 1951, 184). An explication, i.e., a set of principles, should yield the considered judgements about a range of cases. An explication is unsatisfactory only if there are considered judgements about cases, for which it

does not yield any judgements at all, or for which it leads to inconsistent judgements (Rawls, 1951, 185). An explication should be "comprehensive", i.e., it should yield more or less all considered judgements, which should be done "with the greatest possible simplicity and elegance" (Rawls, 1951, 186). Moreover, he relates simplicity to the number of principles used in an explication. Among the tests for the reasonableness of accepting a principle (or a set thereof) as justifiable, Rawls (1951, 188) states its ability to settle existing and new disputed cases, and he relates it explicitly to the virtue of novel prediction for empirical theories.

Rawls also includes virtues which are clearly pragmatic: principles should be applicable to cases yielding judgements, and they should be easy to understand (see also Rawls, 1980, 561). It is interesting to note that Rechnitzer (2022, 116–118) includes fairly similar virtues (practicability, determinacy, broad scope and simplicity) for her detailed application of RE.

Apart from his well-known extension of RE with background theories (see Section 2.2), Daniels draws and defends an analogy between the function of coherence constraints on theory acceptance function in wide reflective equilibrium and science (1979, 257, 273, 279). Simple coherence considerations, e.g., consistency and mere fit between judgements and principles is not enough for their justification (Daniels, 1979, 257). Taking background theories into account may exert more pressure for revision (Daniels, 1980, 86, footnote) and (Daniels, 1979, 258, footnote).

However, background theories are not the only source of revisionary power for RE according to Daniels. The following virtues also figure in Daniels' work on RE. A coherent system of beliefs exhibits fit ("match") between theory and commitments (Daniels, 1996, 2), as well as systematic unity and comprehensiveness (Daniels, 1996, 10). Scope is a virtue of background theories that should reach beyond the commitments (Daniels, 1979, 259). Constraints on theory acceptance in science include considerations of simplicity and parsimony. (Daniels, 1979, 279). He emphasises this point:

> Coherence involves more than mere logical consistency. As in the sciences, for example, we often rely on inference to the best explanation and arguments about plausibility and simplicity to support some of our beliefs in light of others. (Daniels, 1996, 2)

In summary, already Rawls (especially, 1951) and Daniels cover a significant amount of prominent theoretical virtues, insisting on a resemblance between inquiry in science and RE.

Elgin's Elaborations Elgin integrates a treatment of theoretical virtues, in her substantial contributions to elaborating RE. Theoretical virtues already surface in joint work with Goodman (1988)[7], but here, I focus on her work on RE. I opt to present quotations in some length to give an impression of her list of theoretical virtues that go under many different names in her work, e.g., "cognitive values" or "epistemic desiderata". We can extract different functions of theoretical in RE from her work. They provide reasons for revision of commitments working against conservativity, they determine amendments and resolve conflicts, and they foster competition towards maximally tenable systems. Moreover, Elgin relates theoretical virtue to the objective of understanding, and she discusses the consequences of including them in an account of RE. Below, I present these points in more detail.

It is important to note that Elgin does not rely on an explicit distinction between commitments and elements of a theory in her account of RE. In turn she speaks of an "account" or a "system", which comprises commitments to object-level statements about a subject matter, but also commitments to epistemic values, standards, criteria and acceptable methods (Elgin, 2017, 84f). At the outset of inquiry, initially tenable claims form a "motley crew" of elements of different sorts: general and particular statements, idealisations, approximations, specifications, judgements, assessments of value, assertions of fact that reflect our "best estimates of how things stand" (Elgin, 1996, 102). Initial tenability is a weak epistemic achievement, that can easily be lost given that we can provide reasons.

As a result, "initial judgments are not comprehensive; they are apt to be jointly untenable; they may fail to serve the purposes to which they are put or to realize the values we want to uphold." (Elgin, 1996, 106). This is Elgin's starting point for a delicate dialectical process of mutual adjustment towards a system in RE.

A collection of initially tenable commitments, even if curtailed, does not form a system or theory (Elgin, 1996, 103). Systematisation aims at increasing tenability (Elgin, 1996, 110), and it is guided by considerations of consistency, cotenability, relevance, and cogency (Elgin, 1996, 104). Consistency serves as a mandatory virtue during the systematisation of initially tenable commitments towards a theory or system. Otherwise, an inconsistent system implies everything. To have a system that has any implications, and against the

[7] The list of virtues that we can compile from (Goodman and Elgin, 1988, 11–25) is already impressive: consistency, simplicity, uniformity, clarity, relevance, informativeness, rightness relative to a particular purpose, accuracy, scope, entrenchment, appropriateness, precision.

strategy to achieve consistency by collecting elements, which have no bearing on each other, newly added elements are required to be relevant to those already accepted, and if their implications are borne out, they increase the system's cogency.

For a system to be in a state of RE, coherence is required, i.e., that the elements are suitably related in a supportive network, such that each element is "reasonable in light of one another" (Elgin, 1996, 107). But this is not enough, since fictions, for example, may have no independent support except their mutual support, or they may disregard contravening considerations. In addition to coherence, a system in RE needs to be underwritten by independently motivated, initially tenable commitments and maximise tenability. The tether to initially tenable commitments provides the basis for justification *of* the system in contrast to coherence, which justifies elements *in* the system (Elgin, 1996, 107).

Reflective equilibrium as coherence with a consistent and comprehensive class of initially tenable commitments is still problematic. Initially tenable commitments should not be given up without reasons, which are provided by conflicts only, at this point. Thus, in absence of clashes, consistent and cotenable commitments are unrevisable. As a remedy, considering "elegance, breadth, economy, and the like" can spark revisions of initially tenable judgements (Elgin, 1996, 108), for example in view of a "highly plausible, robust and fruitful" or a "powerful" principle. Moreover, these considerations are also directed against conservativity:

> [...] maximizing tenability is not always a matter of minimizing deviation. A system that incorporates a radical hypothesis may be more tenable than its conservative rivals. Even if the hypothesis has no initial tenability of its own, its incorporation might, for example, enable a system to accommodate a higher proportion of our initially tenable commitments; or fruitfully extend beyond its current domain; or avoid ad hoc, implausible, or otherwise untenable assumptions that its competitors have to make. (Elgin, 1996, 109)

Elgin distances herself from the view that knowledge is the resulting epistemic achievement of a system in RE (Elgin, 1996, 122ff). Instead, she promotes understanding as epistemic achievement of RE. In contrast to knowledge, understanding does not apply to facts only, it may be expressed in a non-propositional manner, and it does allow for degrees.

First, Elgin rejects that theoretical virtues have instrumental value by being truth-conducive:

> Tradition has it that truth is our overriding cognitive objective. Even if simplicity, sensitivity, fruitfulness, and the like are genuine goods, their value is supposed to be instrumental, residing in their capacity to promote the discovery of truth. On examination, however, the values in question display little sign of such capacity. Science no doubt favors simplicity. But the simplest theory compatible with the evidence typically has less chance of being true than some of its rivals. (Elgin, 1996, 124)

Moreover, she also rejects the view that theoretical virtues are non-instrumental, subsidiary goods to truth, that "delineate the class of truths a discipline takes for its own." (Elgin, 1996, 125). She rejects this view in light of scientific approximations, idealisations and the like that sacrifice truth for other cognitive values.

Finally, Elgin also considers and rejects an alternative candidate for overriding end: permanent tenability, i.e., tenability that, after some given time, is never lost (Elgin, 1996, 126). However, this threatens to lower the standards for accepting a system in order to secure that we do not have to give them up. Systems "seek a reasonable balance of safety and strength. [...] Cognitive values like informativeness, insightfulness, precision, and predictive power would be forfeit" (Elgin, 1996, 127).

In contrast, theoretical virtues enter the picture as "cognitive values" that foster understanding:

> "Simplicity, sensitivity, explanatory power, and the rest are epistemically creditable not because they are conducive of truth or because they circumscribe a particular class of truths but because they belong with truth to a constellation of cognitive values whose realization promotes the sort of understanding science seeks. (Elgin, 1996, 126)

> We devise a flexible network of cognitive commitments that, through continual readjustments, achieves an understanding of the topic that is on balance reasonable. None of the commitments is absolutely unrevisable. Different potential revisions have different costs and benefits. To decide among potential revisions requires asking what sort of understanding we seek, what resources we

> have to draw on, and what limitations we currently face. We have multiple cognitive desiderata-simplicity, fecundity, elegance, predictive power, and so on. Insofar as is feasible, revisions in our initially tenable commitments should yield an account that satisfies them. (Elgin, 2017, 85)

Elgin also attends to the consequences of including theoretical virtues in RE. Different subject matters or different objectives pursued by inquiry with RE ask for different kinds of understanding, which influence the "constellation" of virtues.

> Other disciplines, having different values and priorities, generate understanding of different kinds. Generality and scope, so central to science, are less important for biography and investigative journalism, where particular actions and events loom larger. But every field of inquiry has its constellation of cognitive values. (Elgin, 1996, 126)

Apart from further virtues provided by science, e.g., empirical and theoretical adequacy, explanatory power and elegance (Elgin, 1996, 139), examples of other kinds of understanding or fields of inquiry having different constellations of cognitive virtues include fictions, or the construction of political systems. The idea, that lists of theoretical virtues differ relative to pragmatic-epistemic objectives relevant to specific domains or problems, surfaces in philosophy of science as well. Kuhn (1977, 362f), for example, suggest changes on his list of theoretical virtues that may be more suitable for engineering (add social utility), or philosophy (remove empirical accuracy).

Next, trade-offs between competing virtues may lead to multiple, equally acceptable outcomes:

> System building is informed by priorities —- second-order commitments about the value of retaining various first-order commitments. Often these determine how conflicts are to be resolved. In empirical science, for example, evidence ordinarily overrides elegance. But our cognitive priorities are neither fine-grained nor regimented. They are unlikely to yield a wellordered ranking of commitments. Competitors in some conflicts thus may have equal claims on our enduring epistemic allegiance. In that case,

> although different resolutions result in divergent systems, the systems that emerge are equally tenable. One system might sacrifice scope to achieve precision, another trade precision for scope. (Elgin, 1996, 134)

And this leads her to adopt a pluralist stance:

> Different accommodations retain different scientific desiderata. Deciding which one to accept involves deciding which features of science we value most and which we are prepared, if reluctantly, to forego.
>
> Pluralism results. The constraints on construction typically are multiply-satisfiable. Where competing considerations are about equal in weight, different trade-offs might reasonably be made, different balances struck. If any system satisfies our standards, several are apt to do so.(Elgin, 1996, 135)

Elgin also states the problem of underdetermination of scientific theories, and proposes a relativisation of truth to system that is still objective because they do not allow for "anything goes" or lead to complete subjectivism (Elgin, 1996, 139–142). For a concise statement of her point we can turn to (Elgin, 1997, 191):

> I have suggested that factual and evaluative sentences are justified in the same way. In both cases, acceptability of an individual sentence derives from its place in a system of considered judgments in reflective equilibrium. Since equilibrium is achieved by adjudication, several systems are apt to be adequate. But since they are the products of different tradeoffs, they are apt to disagree about the acceptability of individual sentences. So relativism follows from pluralism. Something that is right relative to one acceptable system may be wrong relative to another.
>
> Still, the verdicts are objective. For the systems that validate them are themselves justified. The accuracy of such a system is attested by its ability to accommodate antecedent convictions and practices; its adequacy, by its ability to realize our objectives. Several applicable systems may possess these abilities; so several answers to a given question, or several courses of action may be right. But not every system possesses them; so not every answer or action is right. The pluralism and relativism I favor thus do not lead to

the conclusion that anything goes. If many things are right, many more remain wrong.

Kappel's Criticism It is interesting to note that considerations of theoretical virtues go mostly unnoticed by critics of RE, which stands in sharp contrast to the critical attention that considered judgements and background theories receive. An exception to this trend is Kappel (2006), who takes a critical stance towards the prospects of providing *meta-justification* to epistemic desiderata involved in RE, i.e., an explanation of why they are truth-conducive (Kappel, 2006, 134). He presents RE to aim at arriving at a set of moral beliefs that exhibit roughly the following epistemic desiderata (Kappel, 2006, 132f):

(i) Consistency

(ii) Systematicity: a belief set should contain explanatory relations.

(iii) Generality: a belief set should contain general beliefs that cover a larger area rather than a smaller one.

(iv) Simplicity: general explanatory beliefs should be few and simple rather than many and complex.

(v) Intuitive acceptability: moral belief sets (or moral theories) should fit our considered moral judgements.

(vi) Trade-offs between desiderata in order to increase overall consistency, systematicity, generality, simplicity, and intuitive acceptability.

(vii) Dialectical force: other things being equal, we have more epistemic reason to accept sets of beliefs the more they display the relevant epistemic desiderata involved in RE.

Given the usual terminological liberty surrounding virtues, (i) – (iv) are easily recognised as theoretical virtues from philosophy of science. Next, I suppose that we can extract at least two desiderata from (v) "intuitive acceptability". First, there is the hybrid virtue of "fit" between commitments and theory. Second, "intuitive" may hint at the moderate foundationalist claim that moral intuitions or considered judgements should be justified to some extent independently of their inferential relations Kappel (2006, 135). Demanding that commitments should have some positive standing (independent of coherence) is a non-theoretical epistemic desideratum directed at

the commitments. (iv) and (vii) are not theoretical virtues. The former acknowledges trade-offs between (i) – (v). The latter relates the acceptability of a set of beliefs to the degree to which virtues are instantiated. Kappel (2006, 133, footnote) construes epistemic reasons to be about the truth of beliefs.

I am not going to engage with Kappel's argumentation, as he directs his negative outlook at an *epistemic interpretation* of RE in the narrow sense of "epistemic", i.e., a form of justification that aims at the attainment of true and the avoidance of false beliefs (Kappel, 2006, 135).[8] It is important to note that he adapts the idea of meta-justification from BonJour (1985, 9), who develops an account of *empirical* knowledge for which truth is a plausible objective. However, this puts into doubt whether this warrants to transfer the call for meta-justification to RE, as RE includes non-empirical beliefs.

Even steadfast proponents of RE (e.g., Elgin, 1996) reject the truth-conduciveness of theoretical virtues. Instead, there is an elegant escape route, which is also acknowledged by Kappel. If RE is devised to achieve more broadly construed pragmatic-epistemic objectives, e.g., understanding as promoted by (Elgin, 1996), the need for a meta-justification of RE with respect to truth is no longer pressing.

4.3 Configuring Theoretical Virtues

What lessons can we draw from this glance at theoretical virtues in philosophy of science and RE? I think that the overview yields mixed results. On the one hand, the consideration of theoretical virtues is present in many accounts of RE, ranging from the classics to elaborate accounts. For RE as an epistemology and a methodology, the addition of theoretical virtues to RE is attractive. As a driving force behind the process, or as a standard for the evaluation of a state, theoretical virtues exert pressure to systematise beyond mere consistency and fit, provide guidance for revisions, and foster pragmatic-epistemic objectives, such as understanding. Thus, the methodological advice to include theoretical virtues epistemologically founded, and the recognition for trade-offs and the influence of objectives of inquiry is insightful.

On the other hand, I think that the lesson from divergent attempts of systematising the virtues, and longstanding issues leads to a sobering conclusion, at least provisionally. Theoretical virtues are not ready to be fruitfully

[8]For a discussion of epistemic justification and truth in RE, see by a proponent of RE, see Tersman (1993, 94–114).

included into RE, at least if we want to move from the general level of discussion of RE as epistemology or methodology towards being an applicable method. Of course, we can build upon the shared appreciation for theoretical virtues among philosophers of science, and demand for RE that theories do justice to theoretical virtues. However, theoretical virtues also bear the potential to make things worse for RE, as we import their issues as well. Ambiguity among virtues do not help to render RE less vague, and they might even fuel the objections. Trade-offs offer additional sources for divergent outcomes, and conservativity is occasionally proposed as a virtue.

If theoretical virtues resisted being built into RE as an applicable method, then the value of such a methodological advice would be severely limited. This would leave us at the level of plausibility considerations. Is it plausible that considering theoretical virtues in RE helps to address objections and strengthen RE as an account of justification? The stalemate, which I perceive in the discussion about RE, illustrates that plausibility considerations can go either way.

The transfer of theoretical virtues from philosophy of science to RE is impeded further, where the superficial structural parallels between the domains break down. First, the preeminence of truth or empirical adequacy as primary goal of science shapes the presentation, systematisation, and discussion of theoretical virtues. In, contrast, RE is discussed in view of a broad range of aims that are tangential to or even independent of truth. Second, commitments do not have the same epistemic standing as evidence.[9] Consequently, we have to remove irrelevant virtues (e.g., casual adequacy) or provide different interpretations for existing ones (e.g., evidential accuracy). Finally, there may be additional theoretical virtues that are relevant in specific philosophical domains, e.g., for moral theories (Timmons, 2012).

So, much more work is required to get from the epistemological or methodological appreciation of theoretical virtues to their fruitful application in RE as a method. A promising starting point to organise this work, is an idea and the term "configuration", which I take up from (Baumberger and Brun, 2021, 7928). A *configuration* of theoretical virtues consist of *selecting*, *specifying*, *weighting*, and *aggregating* virtues in view of broader pragmatic-epistemic objectives or specific purposes of inquiry.

[9] This is illustrated by the reluctance of scientists to dismiss recalcitrant data. In contrast, initial commitments as "hunches" about a subject matter will or should be much more susceptible to revision if they are in conflict with a theory. For a similar point, see (Arras, 2007, 58)

Selecting and specifying theoretical virtues tackles Kuhn's issue of ambiguity. His second issue, trade-offs, is addressed by weighting and aggregating theoretical virtues. It is the latter two aspects that render a configuration operational for application.

For examples from science of how pragmatic-epistemic objectives can influence the configuration of theoretical virtues, see (Elliott and McKaughan, 2014). They stress the importance to be as explicit as possible about the objectives that are pursued in assessing models, theories or hypotheses (Elliott and McKaughan, 2014, 19). I suppose that the same should hold for the objectives pursued with RE, and it is highlighted by Rechnitzer (2022, 240).

The pragmatic-epistemic objectives or purposes that guide the configuration of theoretical virtues should be clearly stated at some point during inquiry. Do you aim at truth, adequacy, understanding, or applicability? Are the virtues relevant to, e.g., mathematical axiom justification or atmospheric general circulation models? The following questions may help to render a configuration as precise as possible:

(Selection) Which theoretical virtues are relevant to the subject matter, the pragmatic-objective, or the purpose of inquiry? How do the selected virtues contribute to achieving the objective, or serving a purpose?

(Specification) How do the selected virtues relate to each other? Are they necessary requirements or desiderata? Are they used comparatively or categorically? How can they be operationalised? Can comparative virtues be measured?

(Weighting) Are some virtues more important than others? Are some virtues granted strict priority, or do they allow for trade-offs? Can trade-offs be expressed by relative weights?

(Aggregation) Are the measures commensurable? Can the measures and their relative weights be combined, resulting in a degree of overall virtuousness?

Note that it may well be that a configuration of theoretical virtues in view of a pragmatic-epistemic objective cannot be settled before inquiry begins. Instead, configuring theoretical virtues may be part and parcel of a process of equilibration. As part of "second-order" commitments (Elgin, 1996, 134), the configuration of theoretical virtues may undergo change as well, and hence be among the results and not among the prerequisites of a process (see also Rechnitzer, 2022, 32; Baumberger and Brun, 2021, 7929; Elgin, 2017, 89).

Hence, I will focus on virtues that are, in my view, generally relevant to an omnipresent objective of RE: coherence. In view of coherence as a general objective of RE, I will develop a configuration of theoretical virtues that can be integrated fruitfully in RE, inasmuch as they allow to address the conservativity and the no-convergence objections. This serves as a proof of concept and as an illustration for how theoretical virtues can be rendered operational for RE in a precise manner. I do not claim that this is the only or the best way to configure theoretical virtues for RE, but aim to provide a "base" configuration at a very general level that can be adapted or extended to more specific RE contexts later.

I take this to be a worthwhile endeavour, as coherence serves as a stepping stone to various other pragmatic-epistemic objectives of inquiry, such as understanding. Thus focusing on coherence is a first step that leaves the door open to exploit connections to other objectives, or to extend the selection for more specific subject matters, later.

Even though that the involvement of coherence in RE is uncontroversial, spelling out coherence is less so. Indeed, construing coherence as mere consistency and everything fitting together may leave us with a notion of coherence that is too weak to equip RE with justificatory power. Background theories and considered judgements are important ingredients to strengthen RE, and I do not want to miss them as a valuable asset to address objections. However, they may treat only the symptoms of an underlying disease. My diagnosis is that the alleged weakness of RE stems from a weak characterisation of coherence in terms of consistency and everything fitting together. In contrast, my proposal to integrate some theoretical virtues aims at a more substantive notion of coherence, which tackles the problem at its roots. Hence, spelling out a more substantive notion of coherence in terms of virtues may contribute to strengthening RE as an account of justification. This is the task that I will take up in the next chapter.

Part II

Formalisation

Chapter 5

Virtue-Based Coherence in a Deductive Framework

There is widespread appreciation of theoretical virtues in philosophy of science, and elaborate accounts of RE happily adopt them. However, this move may also import the unsettled issues of ambiguity and trade-offs. If we do not want to rely on the authority of prominent figures in philosophy of science, or on the appeal of vague ideas that translate to RE due to structural parallels between scientific inquiry and RE, we need to develop a *configuration* of theoretical virtues that is suitable for an RE setting, i.e., we need to select, specify, aggregate and weigh virtues in view of pragmatic-epistemic objectives.

In the previous chapter, I presented Kappel's critique of RE which revolves around theoretical virtues. He provides a list of desiderata and suggests that

> [...] it may be most appropriate simply to take 'coherence' to label some set of epistemic desiderata much like the ones outlined above, in particular those of consistency, systematicity, generality, and simplicity (Kappel, 2006, 135),

Similarly, virtues figure in Setiya's description of coherence, which he then relates to RE (2012, 26).

> The simplest picture is one of pure coherence: one's ethical beliefs are justified insofar as they belong to a system of beliefs that is simple, powerful, consistent, and explanatorily deep. (Setiya, 2012, 25)

Kappel sets the task of "[s]tating the desiderata more precisely and sorting out their interrelations would be of importance if our aim were to provide a full defence of MRE [method of reflective equilibrium]" (Kappel, 2006, 132).

The aim of this chapter is to take up Kappel's task and show that it can be accomplished. To this purpose, I focus on coherence as an objective that guides the configuring of virtues for RE, and then take the first two steps towards a configuration by selecting and specifying virtues. Coherence, is a general objective of RE that can be related to other pragmatic-epistemic objectives of inquiry, e.g., understanding. Moreover, the contribution of theoretical virtues to coherence is recognised in the RE literature (e.g., Rechnitzer, 2022, 31f). Surely, there remain many other objectives and virtues that have no bearing on coherence, but for now, it useful to provide a "base" configuration at a very general level that can be adapted or extended to more specific RE contexts later.

Very roughly, apart from being consistent with each other, the elements of a coherent system should "hang together" (BonJour, 1985, 93), i.e., they should form a system of mutually supportive elements. On many occasions, the nature of the support relation is taken to be inferential (e.g., BonJour, 1985, 96). For the sake of arriving at a workable configuration, I focus on deductive inference as one aspect of an inferential support relation

I proceed as follows: In Section 5.1, I introduce a propositional framework of deductive inference, and I show that inferential relations on their own fall short of providing a satisfactory characterisation of coherence. Instead of giving up the framework, I suggest in Section 5.2 to impose more structure and include additional virtues in the framework. This results in a more substantial characterisation of coherence that allows to disentangle some complex interrelationships among virtues in Section 5.3. Finally, I discuss the upshots for RE in Section 5.4.

5.1 Coherence and Deductive Inference

5.1.1 A Simple Framework of Deductive Inference

In order to spell out coherence, I adapt the formal framework of *Belief Revision Theory* (BRT, for short). Formal characterisations of coherence have been discussed in BRT since its advent by a seminal paper of Alchourrón, Gärdenfors, and Makinson (1985). The framework consists of a propositional language $\mathfrak{L}$ with usual connectives $\neg, \wedge, \vee, \rightarrow$ and $\leftrightarrow$. Lower-case Roman letters ($p, q, \ldots$) denote atomic sentences. Atomic sentences and their negations are called *literals*. Lower-case Greek letters ($\alpha, \beta, \ldots$) represent formulas over elements of $\mathfrak{L}$, and upper-case Roman letters ($A, B, \ldots$) denote subsets

of $\mathfrak{L}$. Cn is a classical consequence operator for sets of formulas.[1] I use $K \vdash \alpha$ as a notational variant that can be used interchangeably with $\alpha \in Cn(K)$ to indicate that α follows logically from K. Conversely, $K \nvdash \alpha$ signifies $\alpha \notin Cn(K)$. A set of sentences K that is closed under deductive inference ($Cn(K) = K$) is a *belief set*, and non-closed sets are referred to as *belief bases*. The falsum $\bot$ is used as a constant symbol for a contradiction in $\mathfrak{L}$.

For the sake of terminological continuity I will speak of "belief", although "acceptance" of a proposition may be more appropriate, comprising weaker forms than a commitment to truth, e.g., guesses or expectations.[2]

5.1.2 Coherence in Deductively Closed Belief Sets

There are two major ingredients to classically understood coherence. A coherent system is required to be consistent and its elements need to be mutually supportive (e.g., BonJour, 1985, 95).

Consistency, is the absence of contradictions, formally $\bot \notin Cn(K)$. There is widespread agreement in the literature, that consistency, on its own, is not enough for coherence, because a collection of completely unrelated elements is trivially consistent.

How much mutual support, in terms of deductive inferential relations, does coherence demand? The absolute maximum has been formulated by Blanshard (1939, 264), who describes an ideally coherent system, in which every element entails and is entailed by the rest of the system. Formally, the first part would amount to require $\{\alpha\} \vdash \beta$ for all elements, α and β, of a belief system K, which immediately renders all elements of K equivalent. In more recent work, this is still discussed as a state of maximal possible coherence (Mackonis, 2013, 983), especially in probabilistic approaches to coherence such as (Bovens and Hartmann, 2003, 611) or (Fitelson, 2003, 194).

A weaker, but highly influential proposal stems from Ewing (2012(1934), 229), who uses only the second part of Blanshard's characterisation, namely that every element of a coherent system is supported by the rest. This idea to spell out mutual support in terms of residuals is discussed in BRT by Hansson and Olsson (1999, 246):

(Supraclassicality) If $K \vdash \alpha$, then K supports α.

(Residual Support) $K \smallsetminus \{\alpha\}$ supports α for all $\alpha \in K$.[3]

[1] For a formal presentation of the classical consequence operator, see Hansson (1999, 26).

[2] For an introduction of such doxastic concepts, see (Rott, 2001, Ch. 1.1).

[3] The residual $K \smallsetminus \{\alpha\}$ denotes the set-theoretic removal (subtraction) of α from K, for example, $\{\alpha, \beta, \gamma\} \smallsetminus \{\alpha\} = \{\beta, \gamma\}$.

(Coherence) If K is consistent and satisfies (Residual Support), then K is coherent.

It is an important result of (Hansson and Olsson, 1999) that deductively closed belief sets collapse so-defined coherence into consistency. Consequently, spelling out the mutual support of coherent systems as deductive inferences from residuals does not work out for closed belief sets. Hansson and Olsson (1999) trace this failure to the introduction of irrelevant deductive relations by closing a set under logical consequences. If $\alpha \in K$, and if K is closed under logical consequences, i.e., $K = Cn(K)$, then we also have $\beta \rightarrow \alpha \in K$ and $\neg\beta \rightarrow \alpha \in K$. This is not altered by removing α from K, and the tautology $\beta \vee \neg\beta$ is element of any belief set. Moreover, $\beta \rightarrow \alpha$, $\neg\beta \rightarrow \alpha$, and $\beta \vee \neg\beta$ jointly entail α. Consequently, $K \smallsetminus \{\alpha\} \vdash \alpha$.

5.1.3 Coherence in Non-Closed Belief Bases

Hansson and Olsson (1999) recommend to consider non-closed sets of sentences, so called *belief bases* as a more discriminative approach. A belief base approach allows to distinguish between non-derived (independently held) and merely derived beliefs (Hansson and Olsson, 1999, 259–261). They suggest, that a belief base representing a belief state should consist exactly of those non-derived beliefs an agent holds independently. Coherence related properties are then to be examined for belief bases, e.g., the list of coherence criteria of BonJour (1985), but they do not develop a full-fledged account of coherence for belief bases.

In a first step towards developing such an account, we may ask, whether mutual support by deductive inference from residuals is a viable approach for belief bases. As it turns out, the belief base approach combined with the exclusive reliance on deductive inference is also beset with problems that call for refinements. A referee pointed out to Hansson and Olsson (1999, 264, footnote 40) that (Residual Support) combined with deductive inferential relations is rarely satisfied. Take for example the following sentences (the example is adapted from (Hansson, 2006, 96):

p: Bob is a Catholic.

q: Bob is ordained

$\neg r$: Bob is not married.

In this example, $K = \{p, q, \neg r\}$ does not satisfy (Residual Support) due to the complete lack of inferential relations between its elements. Clearly, we

could try to alleviate the situation by extending K with plausible inferential relations, e.g., $K' = K \cup \{(p \wedge q) \rightarrow \neg r\}$ Still, (Residual Support) would not be satisfied due to $K' \smallsetminus \{p\} \nvdash p$.

In contrast to the suspicion that (Residual Support) is hard to come by, it seems to me that the belief base approach and deductive inferential relations allow for what I call "syntactical trickery", which effortlessly renders consistent belief bases coherent. Consider the following strategy: Let $\bigwedge K$ denote the conjunction of all elements in K, so $\bigwedge K = p \wedge q \wedge \neg r$ in the example from above. It is easy to check that extending K by $\bigwedge K$ satisfies (Residual Support):

$$(K \cup \bigwedge K) \smallsetminus \{\alpha\} \vdash \alpha$$

for all $\alpha \in K \cup \bigwedge K$. This holds in general as well. If K is finite and consistent, $K \cup \bigwedge K$ satisfies (Residual Support), and thus, is coherent. This means that it is very easy to render a finite and consistent belief base coherent. However, it seems strange to me, that the presence of a single element that conjoins all sentences should render a set of mutually unsupportive elements coherent. Given the belief base approach, one can of course question whether such a conjunction has independent standing, and thus be an element of a belief base. The construction of above example gives impression that the conjunction is merely derived. Unfortunately, there is no criterion that would disqualify a conjunction as basic, non-derived belief. As it stands, every set of sentences is a belief base (Hansson, 1996, 200; 1999, 18).

Another way to put the matter is this: The presence of $\bigwedge K$ in $K \cup \bigwedge K$ is almost redundant. Here, "almost" is important. I do not want to say that closing operations that arise from classical rules for conjunction introduction and elimination have no bearing on coherence, they are inferential relations after all. Clearly, if an agent accepts α and β, they should also accept $\alpha \wedge \beta$, and vice versa. However, such operations should not do the major work towards establishing coherence, because they exploit the fact that set-theoretic subtraction of α from a set K is too weak to remove redundancies introduced by sentences containing α as a conjunct.[4] In this context, Bartelborth (1999, 212) speaks of *deductive redundancy*, and concludes that it renders the exclusive reliance on deductive inference overly simplistic for coherence.

The lesson that we can draw from this goes as follows: If we construe coherence as a consistent set of of mutually supportive elements, where the

[4] BRT offers stronger operations than set-theoretic subtraction, e.g., contraction $\div$, which is guaranteed to be successful, i.e. $\alpha \notin Cn(K \div \alpha)$ as long as α is not a tautology. However, contraction is far too strong to figure sensibly in (Residual Support).

support relation is understood purely as deductive inference from residuals, then it collapses into consistency for deductively closed sets, and it can trivially be established for belief bases.

We could try to escape this situation by including non-deductive inference as additional support relations, which would render the problem of characterising coherence substantially more complex or very general. For a formal treatment of a general support relation that includes inferential, explanatory, justificatory, or probabilistic relations, see (Hansson, 2007, 291).

Still, I think that we can do more to capture coherence in a simple framework by deductive inference complemented with additional virtues. Multicriteria approaches to coherence have been proposed earlier by BonJour (1985, 95–99), or Thagard (2000, 53). Arguably, my proposal does not mark a complete departure from their work.

5.2 Virtue-Based Coherence

5.2.1 More Structure for Belief Bases: Literal and Inferential Beliefs

If we want to get more out of belief bases with deductive inference, we have to invest something. Here, the investment will be assumptions that allow to equip belief bases with more structure.

Having RE in mind, we could just impose the distinction between commitments, systematic elements of theories and background on belief bases and see where this gets us. I would like to pursue a more subtle approach that sets out from a distinction between "literal" and "inferential beliefs" in a belief base.

A *literal* belief is represented by a literal of the formal language $\mathcal{L}$, i.e., an atom or a negation thereof. Let us assume that the set of literals $\mathfrak{L}$ consists of finitely many sentences that are relevant to a subject matter. We reduce multiply negated elements and identify $\neg\neg\lambda$ with λ. Disjunctions of literals represent *inferential* beliefs. An inferential belief is supposed to capture the inferential relations that obtain between the elements of a deductively valid argument accepted by the agent.[5]

[5] A disjunction of literals $\lambda_1 \vee \cdots \vee \lambda_n$ is equivalent to $(\neg\lambda_1 \wedge \ldots \neg\lambda_{n-1}) \rightarrow \lambda_n$ (and any permutation of literals therein). The deduction theorem of classical propositional logic entails $\{\neg\lambda_1, \ldots, \neg\lambda_{n-1}\} \vdash \lambda_n$, that is, λ_n is the conclusion of a deductively valid argument with premises $\neg\lambda_1, \ldots, \neg\lambda_{n-1}$.

Assume that a belief base consists of literal and inferential beliefs. Every set of sentences (belief base) can be brought into this special form but the correspondence is one-to-many, unfortunately. For example, if we would try to get there by means of collecting the elements of a belief base into a conjunctive normal form. A formula in conjunctive normal form is a conjunction of disjunctions of literals, and it could be split easily into literal and inferential beliefs. However, while this can be done for every formula, there are multiple, equivalent conjunctive normal forms for the same formula, which would result in different partitionings of a belief base into literal and inferential beliefs.

In view of this, we need to assume that the work of organising the belief base has been completed. Let me sketch how this work could be done in a more promising way than conjunctive normal forms. Rather than starting from belief bases which represent beliefs in a syntactical form, we go back to the basic idea of BRT that set of sentences model *belief states*. How do we assign a set of sentences to a belief state? Hansson (1999, 9) proposes that a sentence p is an element of a belief set if and only if the question "Do you believe that p?" is answered affirmatively by a system in a specific epistemic state.[6]

I suggest we can proceed similarly for belief bases. Assume that the formal language $\mathcal{L}$ is given. For every literal λ from $\mathfrak{L}$ we ask an agent in an epistemic state:

Do you accept λ independently of other beliefs that you hold?

The answers to these questions give rise to a set L of literal beliefs that an agent holds in their current epistemic state. Note that an agent is able to abstent from accepting or rejecting (i.e., accepting the negation of) a literal belief resulting in suspension of belief. The demand for independently held beliefs is directed against situations, in which the agent holds a literal belief λ only because they derive it from other beliefs, e.g., from $\beta \vee \lambda$, $\neg\beta \vee \lambda$ and $\beta \vee \neg\beta$. It is an intricate task to spell out what it means to hold a belief independently. Hansson (1999, 21f) describes the elements of a base as "self-sustained" beliefs, that "are worth retaining for their own sake", and he provides examples of beliefs that are based on memories, or on beliefs for which we lost track of their derivation.

Next, we can ask for any combination of literals from $\mathfrak{L}$ of length k:

[6] This presuppose a very highly idealised system (e.g., a database on a computer), and ascribing beliefs to human beings would be part of a much more intricate process that takes many factors (beside answers to explicit questions) into account (Rott, 2001, 10f).

> Do you accept $\lambda_1 \vee \cdots \vee \lambda_k$ independently of other beliefs that you hold?

Again, asking for independently held beliefs prevents cases in which an agent merely derives inferential relations from other beliefs, e.g., $\lambda \vee \alpha$ from a literal λ. The answers can be collected into a set I of inferential beliefs.

Note that the organisation of belief bases prevents the "syntactical trickery" from above. One can no longer establish mutual support from residuals by introducing a conjunction of all elements. Take for example

$$K = \{p, q, r, (p \wedge q) \rightarrow \neg r, p \wedge q \wedge \neg r \wedge ((p \wedge q) \rightarrow \neg r)\},$$

which satisfies (Residual Support) due to including a long conjunction. After imposing more structure on this belief base, it may look like this:

$$K' = L \cup I = \{p, q, \neg r, \} \cup \{\neg p \vee \neg q \vee \neg r\},$$

which no longer involves conjunctions that could easily establish (Residual Support).

Note that I take the idealised question answering to aim at an *initial* belief base. At this point, a belief base does not need to satisfy any constraints of rationality. For example, a belief base may be inconsistent, or it may not include all logical consequences, e.g., q, even though an agent accepts p and $\neg p \vee q$ independently of other beliefs. Afterwards, coherence considerations advise to render a belief base consistent, and to make merely derived beliefs explicit by including them, or revise or give up on initially independently held beliefs.

5.2.2 Account

Despite representing the epistemic state of agent with a more structured belief base, the aim of coherence remains the same, i.e., to establish a network of mutually supportive elements. Given a belief base separated into literals L and inferential beliefs I, I suppose that relevant support relations for coherence are inferential relations which occur among literals in L given the inferential background I. L, on its own, does not exhibit any inferential relations, as it consist of literals. Inferential relations between L and I are again of lesser interest from the perspective of coherence due to disjunction introduction. The fact that $\neg r$ is in L suffices to infer $\neg p \vee \neg q \vee \neg r$ or any other disjunction having $\neg r$ as a conjunct.

We can reframe (Residual Support) for a belief base $B = L \cup I$ by requiring every literal in L is supported by the rest of the literals given inferential relations I. Formally,

(Residual Support') $L \setminus \{\lambda\} \cup I \vdash \lambda$ for all $\lambda \in L$.

Given a set of literals L, we can try to characterise a set of inferential beliefs I such that (Residual Support') will be satisfied. The following proposition relates (Residual Support') to the presence of inferential beliefs of a specific form among the consequences of I:

Proposition 1. *Let $L = \{\lambda_1, \ldots, \lambda_n\}$ be a set of literals. $B = L \cup I$ satisfies (Residual Support') if and only if*

$$\lambda_k \vee \bigvee_{\substack{i=1 \\ i \neq k}}^{n} \neg\lambda_i \in Cn(I)$$

for all $k \in \{1, \ldots, n\}$.

Proof. See Appendix B.1. □

The condition on the right-hand side of the equivalence in Proposition 1 can be met, which is to say that there belief bases that satisfy (Residual Support'), as the following example illustrates:

$$L = \{\alpha, \beta, \gamma\}$$
$$I = \left\{ \begin{array}{l} \neg\alpha \vee \neg\beta \vee \gamma, \\ \neg\alpha \vee \beta \vee \neg\gamma, \\ \alpha \vee \neg\beta \vee \neg\gamma \end{array} \right\}$$

The formulas of I are equivalent to $(\alpha \wedge \beta) \rightarrow \gamma$, $(\alpha \wedge \gamma) \rightarrow \beta$, and $(\beta \wedge \gamma) \rightarrow \alpha$. So, any two literals of L imply the third.

The question remains whether this condition, or equivalently, (Residual Support'), is met in interesting or somewhat realistic cases. Take again the example from earlier about Bob, an unmarried ($\neg r$), Catholic (p) priest (q), and $B = L \cup I$, where

$$L = \{p, q, \neg r\}$$
$$I = \{\neg p \vee \neg q \vee \neg r\}$$

B does not satisfy (Residual Support') because

$$\{p, \neg r\} \cup I \nvdash q$$
$$\{q, \neg r\} \cup I \nvdash p.$$

One could try to alleviate the situation by extending I with $\neg p \vee r \vee q$ ("If Bob is an unmarried Catholic, then he is ordained.") and $\neg q \vee r \vee p$ ("If Bob is an unmarried priest, then he is Catholic."). However, it seems to me that it is rather hard to imagine plausible scenarios, in which those inferential beliefs can be held independently. This stands in sharp contrast to $\neg p \vee \neg q \vee \neg r$ ("If Bob is a catholic priest, then he is unmarried.") being an element of I, which could be held independently due to being recalled from a memory about Catholicism, for example. Note that other kinds of support relations, e.g., probabilistic ones, manage such cases better. For example, it is plausible that Bob being an unmarried Catholic makes it more likely that he is ordained. So, given the sole reliance on deductive inferences, (Residual Support') may still be too demanding as an requirement of coherence.

Another option for improving the situation would be to replace "for all $\lambda \in L$" in (Residual Support') by "for some $\lambda \in L$" and allow for degrees of residual support. To do so amounts to acknowledge functional differences among elements in L. Some subsets of L follow deductively from others given the inferential background I. In our example, $\{p, q\}$ allows to infer $\neg r$ given background I. Other subsets, e.g., $\{q, \neg r\}$ do not have this property of implying additional elements of L besides themselves. Thus, there are elements, which jointly serve to *axiomatise* (parts of) L. Let us call such subsets of L *theories*. Note that there are no restrictions on which elements can be members of a theory, e.g., on the basis of their generality. At this point, any subset $T \subseteq L$ qualifies as a theory. Subsequently, we will be preoccupied with identifying desirable features that make some subsets "better" than others.

I call the support relation that holds between a theory and the other elements of L *account* (given background I). A theory accounts for elements in L, but not vice versa, this stresses the asymmetric nature of account in contrast to "fit" or "hanging together".[7] A theory T can account for more or

[7] It is a different question, for which relations support would "flow" back from the accounted for elements in L to the theory T that axiomatises them. Inductive or probabilistic relations could serve this purpose. However, it goes beyond the scope of this project to examine possible extension of the present account.

less elements in L, and hence account comes in degrees. Formally, we can characterise account as follows:

(Account) The degree to which T accounts for L is proportional to the number of $\lambda \in L$, such that $T \cup I \vdash \lambda$.

(Account) is a gradual notion but it can be strengthened into a categorical notion if every literal is accounted for:

(Full Account) $T \cup I \vdash \lambda$ for all $\lambda \in L$.

There is still more that we might demand. The logical consequences of T given I may comprise additional literals that are not in L, e.g., q for $T = \{p\}$ and $I = \{\neg p \vee q\}$. If a theory T accounts for all literals of L, and only for those, we have *full and exclusive account* (FEA). Let me introduce notation to ease the formalisation of (FEA): Assume that a set of inferential beliefs I is given, and recall that $\mathfrak{L}$ is the set of literals of our formal language. Let $\overline{T}$ denote the logical consequences of T (given I) restricted to $\mathfrak{L}$. Formally,

$$\overline{T} = Cn(T \cup I) \cap \mathfrak{L}.$$

Given this notation, we can express full and exclusive account concisely:

(Full and Exclusive Account) $\overline{T} = L$

Similarly, (Full Account) can be expressed as $L \subseteq \overline{T}$. Consequently, (FEA) implies (Full Account).

Note that (Full Account) seems very similar to (Residual Support').[8] The former requires that every element is accounted for by a theory, the latter demands that every element is supported by the rest. Thus, the difference is this: For (Full Account) a single subset, the theory, does all the work. In contrast, for (Residual Support'), the multiple subsets, the "rest", vary for every element in question.

Another way to portray the difference is the following: There is a trivial inferential relation $\{\lambda\} \cup I \vdash \lambda$ because the tautology $\lambda \vee \neg\lambda$ is always in $Cn(I)$. In this sense, every element is able to support itself. The more demanding (Residual Support') prevents the exploitation of this trivial inferential relation by subtracting λ set-theoretically from the premises of its

[8] Despite their similarity, (Full Account) and (Residual Support') are mostly independent of each other: (Full Account) does not entail (Residual Support'). If (Residual Support') is satisfied, (Full Account) may still only be established by the trivial theory that repeats every literal (see the example after Proposition 1).

derivation in $(L \setminus \{\lambda\}) \cup I \vdash \lambda$. (Residual Support') requires that there is no element in L that is supported only by itself. In contrast, (Account) allows for elements which are accounted for in this trivial fashion. In the extreme, (Full Account) could be established easily by simply repeating L as theory. As it stands, the account relation on its own falls short of providing a substantive characterisation of coherence.

I think that the subsequent virtues can deal with the issue of establishing (Account) by questionable means. Hence, I pursue the idea of (Account) rather than (Residual Support'), complement it with other virtues, and see whether they can work together to restrict problematic cases such as repeating literals in the theory that only account for themselves.

(Account) is related to coherence in a straightforward manner as it captures inferential relations among literals in a belief system. The more elements of L are accounted for by a theory $T \subseteq L$ (given I) the more coherent is $B = L \cup I$. Note that this degree has to be understood relative to $\overline{T}$ or $\mathfrak{L}$, and this applies to subsequent gradual virtues, too. Otherwise, bigger sets of beliefs will tend to be more coherent just because they exhibit more inferential relations.[9]

Concerning account in relation to theoretical virtues in philosophy of science, I take account to be faintly similar to "accuracy", e.g., as presented by Kuhn (1977, 357):

> First, a theory should be accurate: within its domain, that is, consequences deducible from a theory should be in demonstrated agreement with results of existing experiments and observations.

It is not my intention to portray account involving beliefs with empirical content, but I would like to emphasise the aspect of agreement. A theory is accurate to literal beliefs to the extent that they are in agreement with the consequences of the theory given inferential relations, and this corresponds to the relation of account. Moreover, every element trivially agrees with itself. So, if we aim for capturing agreement with (Account), we might be reluctant to remove literals preemptively as (Residual Support') advises.

In the literature about RE, the idea of agreement between the set of commitments and the theory has been promoted from the early groundwork of Goodman (1983(1955), 64) to the most recent application (Rechnitzer, 2022, 24f). For a system under construction in an equilibration process, Elgin (1996,

[9] For a critical remark on BonJour's criterion that the coherence of a system depends on the absolute number of inferential relations (BonJour, 1985, 98), see (Olsson, 2021).

107) advises "to test the construction for accuracy by seeing whether it reflects (closely enough) the initially tenable judgments we began with".

5.2.3 Consistency

The virtue of consistency is omnipresent in philosophy of science, and its relation to coherence is quite obvious. Consistency is often stated as a necessary but insufficient condition for coherence.

In the present setting, we can motivate the virtue of consistency in view of (Account), as inconsistent theories immediately establish (Full Account). The "principle of explosion" (*ex falso quodlibet*) from classical propositional logic postulates that anything can be inferred from a contradiction. Hence, an inconsistent theory would account for everything. Clearly, this is not the kind of inferential relation we want to be exploited for account. This motivates to include the virtue of consistency as a requirement for theories. Minimally, this includes the *internal* consistency of a theory T ($\bot \notin Cn(T)$), i.e., it must not entail flat contradictions, e.g., λ and $\neg\lambda$. Moreover, T needs to be consistent with the inferential background I as well. If it is assumed that some inferential beliefs arise from other theories in the background, then the we could speak of *external* consistency of T given I. Formally, internal and external consistency are covered by requiring

(Consistency) $\bot \notin Cn(T \cup I)$.[10]

Note that it is a welcome advantage of the present belief base approach that we do not need to require that the entire initial belief base $B = L \cup I$, is consistent, i.e., $\bot \notin Cn(L \cup I)$, before we can operate on it meaningfully.[11] This is not to say that consistency is not an ideal of coherence. (Full Account) cannot be achieved by a consistent theory if B is inconsistent. Thus, the present setting allows to see whether or in which circumstances the consistency requirement for theory and the desideratum of account "push" consistency into the entire belief base.

[10] As a consequence, I, the set of inferential beliefs, needs to be consistent as well ($\bot \notin Cn(I)$). This amounts to the requirement, that inferential relations of I need to be satisfiable, i.e., there is a truth-value assignment such that every formula in I evaluates to "true". There are rather peculiar arrangements of inferential beliefs, such as $I = \{p \vee q, p \vee \neg q, \neg \vee q, \neg p \vee \neg q\}$, that are unsatisfiable.

[11] This is an advantage over deductively closed belief sets, which comprise the entire language $\mathcal{L}$ in case of inconsistencies. Belief bases come with more expressive power, as they allow to distinguish different inconsistent bases.

5.2.4 Simplicity

A plausible countermeasure to establish account with elements that account only for themselves, or by enlisting all literals as theory, is the virtue of simplicity. Simplicity is probably the most widely discussed theoretical virtue in philosophy of science. We can distinguish between *ontological* and *syntactical* simplicity of a theory, which comes down to the demand of postulating fewer entities or, respectively, fewer principles. Clearly, we can only consider syntactical simplicity in the present setting, as propositional variables do not provide insight into what they postulate ontologically.[12] The current representation of a belief state as a belief base consisting of literal and inferential beliefs prevents the conjoining of literals into a single "theory sentence". Hence, the size of the theory is a suitable approximation of its syntactical complexity, and conversely:

(Syntactic Simplicity) The degree of syntactic simplicity of T is inversely proportional to $|T|$.

Now, achieving a higher degree of account by including otherwise inferentially inert literals comes at the cost of lower syntactical simplicity.

Note that the syntactic simplicity of a theory is defined in absolute terms. However, one might wonder, whether we should not better relativise simplicity to the number of literals among the consequences of a theory. After all, a theory may still be considered "simple" relative to a great amount consequences that we can derive from the theory. However, I will soon introduce a different notion, that can be defined in terms of syntactic simplicity and the subsequent virtue of scope, rather than opting for relativising simplicity.

How does simplicity relate to coherence? This is not obvious because simplicity does not establish inferential relations, which are essential to coherence. However, simplicity works against adopting inferentially unrelated subsystems, which BonJour (1985, 97) takes to be detrimental to coherence. This is underwritten by Kuhn (1977, 357), demanding that a theory "should be simple, bringing order to phenomena that in its absence would be individually isolated and, as a set, confused". In the present framework, inferentially inert literals exemplify unrelated subsystem as they do not exploit inferential relations to other elements. Having to account for literals in inferentially unrelated subsystems will result in syntactically complex theories.

[12] Obviously, the propositional framework also does not allow to examine the syntax of propositional variables which renders "syntactic simplicity" a slight misnomer. Nonetheless, I continue to use this fairly standard piece of terminology. See, for example, Baker (2022) or Schindler (2018).

5.2.5 Scope

Account involves the deductive closure of a theory T (given background I) $Cn(T \cup I)$, in relation to literals in L. Recall that $Cn(T \cup I)$ restricted to all literals $\mathfrak{L}$ of the language is denoted by $\overline{T}$. This allows to consider the deductive closure of a theory independent of a set of literals L.

(Scope) The scope of T is proportional to $|\overline{T}|$.

Scope is recognised as a theoretical virtue, e.g., (Kuhn, 1977, 357), or under the name of "strength" by Lewis (1973, 73), but is it relevant to coherence? Let me present a series of motivations for scope that are mostly independent of coherence as they may come to mind first.

In philosophy of science, the idea seems roughly to be that broader scientific theories have more potential for novel prediction ("fertility", "fecundity"). However, novel prediction does not figure as a sought-after goal of coherence. In RE, however, we might pursue the pragmatic objective of being able to handle new cases, which would be facilitated by scope. If aim for a more epistemic spin of this idea, we might turn to the aspect of stability present in state of (provisional) equilibrium. Broad scope might increase the stability of a state as it is able to accommodate otherwise disrupting information. However, I am not going to speculate about these ideas further, as they relate to other pragmatic-epistemic objectives than coherence that we might pursue with RE

Next, scope might be of value in a dynamical setting of belief change, as it might provide additional guidance that goes beyond account, consistency, and simplicity. If the scope of a theory goes beyond the literals of a belief base $B = L \cup I$, a sensible adjustment would be to include the missing literals. In this case the updated belief system would be more coherent than its predecessor due to more accounting relations.

However, there is also a notable overlap of account and scope bearing the risk of double counting. Elements of $\overline{T}$ both contribute to account for literals in L as well as having broad scope. A broadly scoped theory has an advantage to perform well according to account. Thus, we may also ask, whether scope is not redundant for coherence in view of account.

I think we can make some points that speak in favour of considering scope for coherence in a static setting, nonetheless. First, it is helpful to distinguish scope and account in terms of *potential* and *actualisation*. Scope concerns potential inferential relations between literals independent of whether they are

in L. Account is actualised scope in a set of literals L. Scope contributes indirectly to the main objective of coherence to have relatively many inferential relationships among the elements of a system.

But is there also a direct motivation to consider scope for coherence in a static setting? I think that there is. Assume that a consistent theory T ($T \subseteq L$ for a belief base $B = L \cup I$) is quite simple, and T accounts for many elements in L, but T has low scope relative to $\mathfrak{L}$. What does that mean? Recall that $\mathfrak{L}$ consist of literals that are relevant to a subject matter. If T has low scope relative to $\mathfrak{L}$, T fails to cover substantial parts of the subject matter. If the objective is to have a coherent set of beliefs about a subject matter, then low scope with respect to the subject matter indicates, that the objective has not been achieved. In this case, an agent will have "blind spots" with respect to other pursued pragmatic-epistemic objectives that build upon coherence. For example, there may be parts of the subject matter, for which an agent lacks understanding, or for which they do not have beliefs (they are not in L), yet possibly justified beliefs (they are not accounted for by T).

The relationship between scope and account remains intricate, even in view of the potential-actual- distinction. Scope is limited by account on the low end. A theory cannot have lower scope than what it accounts for. At the other end, scope can go beyond account if there are literals μ of the formal language, for which $\mu \notin L$ but $Cn(T \cup I) \vdash \mu$. These elements are not captured by account, and this motivate the consideration of scope. This can occur, if the agent has note made up their mind about all literals of a subject matter, i.e. if they neither accept nor reject some literals. In setting, where L contains a literal for every atomic sentence in the formal language, indicating that the agent either accepts or rejects every atomic sentence that is relevant to a subject matter, scope is covered by account. In this case, (Full Account) implies maximal scope. Note that the scope of a theory going beyond the literals in L is incompatible with full and exclusive account (FEA).[13]

Concerning RE, it seems to me that we can understand Rawls' demand for a comprehensive explication of considered judgements (1951), or Daniel's inclusion of background theories (1979) to be directed at establishing sufficiently broadly scoped coherent systems of beliefs.

[13]In a dynamic setting, at the initial stages of inquiry, aiming for scope helps to prevent that the theory is just designed to fit the initially held literals to a tee. It seems plausible to me, that at later stages during inquiry, (FEA) should receive more weight than scope in light of coherence as a goal. Otherwise, the full potential of scope is never tapped.

5.3 Additional Virtues

So far, my proposal to spell out coherence in a simple framework involves the virtues of consistency, account, simplicity and scope. This may look rather meager in view of the sheer endless lists of theoretical virtues. Let me illustrate how the presented virtues relate to two other, commonly discussed virtues: *unification* and *non-ad-hocness*. This shows that the proposed framework is able to accommodate additional virtues with simple means disentangling their complex interrelationships. Moreover, I suppose that both unification and non-ad-hocness relate to coherence as well.

5.3.1 Unification

Unification receives a lot of attention in philosophy of science, resulting in various proposals that I cannot present in detail. For the present purpose, it is enough to get a rough idea that allows to be taken up in the present framework. Prominently, Friedman describes unification in science as follows:

> I claim that this [unifying effect] is the crucial property of scientific theories we are looking for; this is the essence of scientific explanation - science increases our understanding of the world by reducing the total number of independent phenomena that we have to accept as ultimate or given. (Friedman, 1974, 15)

So, Friedman, and with him many others, regard unification as a theoretical virtue in relation to explanation and understanding. Subsequently, Kitcher (1976) revealed flaws in Friedman's formal proposal, and he takes the following as a starting point for his own account of unification. He construes unification to aim at "the best tradeoff between minimizing the number of premises used and maximizing the number of conclusions obtained" (Kitcher, 1989, 431).

I suppose this is suitable to establish a relation to theoretical virtues in the present framework as well. "Minimizing the number of premises" translates roughly to the virtue of syntactical simplicity, "maximizing the number of conclusions" to scope in the present framework, and there is need to trade them off against each other. This is also inline with Hempel's characterisation of "systematic power":

> Some theories seem powerful in the sense of permitting the derivation of many data from a small amount of initial information;

> others seem less powerful, demanding comparatively more initial data, or yielding fewer results. (Hempel, 1965, 278)

Hempel presents a formally precise definition of the systematic power of a theory "reflected in the ratio of the amount of information derivable by means of T to the amount of initial information required for that derivation" (Hempel, 1965, 279).

The idea that we can import from Kitcher's and Hempel's characterisations to the present framework is straightforward: Simplicity and scope can join forces to yield unification. A simple theory that has broad scope exhibits *unifying potential*. A theory with unifying potential *actually unifies* (provides unification for) the elements it accounts for.[14] Unifying potential relates two virtues, simplicity and scope, actual unification involves account as a third element. As simplicity, scope and account come in degrees, the instantiating of unifying potential and actual unification of a belief system will be gradual notions as well.

How much simplicity can be given up to broaden scope without hampering unifying potential is subject to a trade-off. If, for example, simplicity is more important than scope, at small reduction of simplicity needs to be accompanied by a substantial gain in scope. The details of such trade-offs will depend on the subject matter at hand, as well as epistemic-pragmatic objectives or purposes of inquiry that go beyond coherence.

Concerning the distinction between unifying potential and actual unification, we have to note that it is rarely made explicit in philosophy of science with some exceptions, e.g., between predictive and explanatory power (Hempel, 1965, 278), or between theoretical unification and unifying power (Schindler, 2018, 12). This may be due to the fact that accuracy is considered to be the most decisive virtue (Kuhn, 1977), or granted lexicographic priority over other virtues, leaving us to consider actual unification only.

I think, that present construal of unification as a suitable trade-off between simplicity and scope is helpful to asses another construal of the relationship between unification and simplicity in philosophy of science, which is different from the present proposal. The idea allows to be expressed in short slogans, such as

[14] Schindler (2018, 173–183) discusses the Glashow–Weinberg–Salam model of electroweak interaction as an example of a theory with unifying potential lacking empirical adequacy at the moment of its advent.

> Unification favors hypotheses that explain more facts with same resources. Simplicity favors hypotheses that explain same facts with fewer resources (Mackonis, 2013, 990)

or

> A theory that exhibits simplicity explains the same facts as rival theories, but with less theoretical content. A unified theory, however, is one that explains more kinds of facts than rival theories with the same amount of theoretical content. (Keas, 2018, 2775 in reference to Thagard, 1978)

I am not too happy about how they relate unification and simplicity, because what they call "simplicity" is just another aspect of unification.[15] Moreover, formulations that include "same" or other ceteris paribus clauses do not rule out that there may be trade-offs, but they are of no help in scenarios, where things may not be that "well-behaved", i.e., if other things are not equal. If rival theories do not have the same amount of theoretical content, they are not comparable with respect to unification, even if they differ drastically in how broadly scoped they are. Similarly, rival theories are incomparable according to simplicity, if they do not have equally broad scope. As simplicity and scope tend to pull in different directions, overall comparisons may rarely obtainable.

Note that unification as a suitable trade-off between syntactic simplicity and scope captures the intuition behind the ceteris paribus clauses from above while evading their problems. It is faithful to the idea that if either syntactic simplicity or scope is equal, improving the other virtue increases unifying potential (or unification in case of account). If two theories are equally simple, the one with broader scope has more unifying potential. In turn, for two theories with equal scope, the one that is more simple, exhibits more unifying potential.

The relations between unification and coherence are abundant in the literature. Schurz (1999, 98) claims that "coherence minus circularity equals unification", and (Bartelborth, 1999) counts explanation as unification among the constituents of coherence. Mackonis (2013) explicates coherence in terms of explanatory virtues, including unification.

[15] To be fair, Mackonis (2013, 991) notices that the kind of simplicity described in his quote above falls under unification.

5.3.2 Non-Ad-Hocness

In the present framework, unification, as interplay between account, simplicity and scope, contributes to the goal of coherence by promoting inferential relations between literals. This can be exemplified by a final virtue of this section: non-ad-hocness. Again, philosophy of science extensively covers the topic, resulting in various accounts and controversy about what exactly constitutes the epistemic deficiency of ad-hocness. This results in multiple proposals for bearers of ad-hocness. Hypotheses as elements of theories, theories or even series of theories have been deemed ad hoc. In the present setting, I attempt to characterise ad hoc elements in a theory. Consequently, a theory will be more or less ad hoc depending on the share of ad hoc elements.

Commonly, ad-hocness is presented in a dynamical setting: ad hoc elements are *introduced* to a theory in a specific situation, e.g., in face of recalcitrant data, and for a specific purpose, e.g., to save it from refutation. This stands in contrast to the present focus on the static aspect of a theory's virtuousness. Nonetheless, I suggest that we can characterise the functional role of ad hoc elements in terms of virtues independently of their introduction to a theory.

In order to get an idea of the role of ad hoc elements, we can have a look at the literature. Classically, Popper (1959b, 62) states that introducing auxiliary hypotheses ad hoc to a theory results in a decrease in falsifiability and testability, and he allows for degrees of ad-hocness that are inversely related to degrees of testability and significance (Popper, 1959a, 50). Lakatos (1978, 68), takes ad-hocness to be a symptom of "degenerating research programmes", i.e., if a theory in a series of theories does not have excess empirical content of which some is corroborated (Lakatos, 1978, 33f). More precise notions of ad-hocness have been developed by Leplin (1975, 336–337), who provides an elaborate list of individually necessary and jointly sufficient conditions of ad-hocness, and by Grünbaum (1976, 333–337), who proposes three definitions of ad-hocness with ascending logical strength.

I cannot do justice to the variety and depth of this debate, but aim to embed a simple idea in the present framework. In a paper with a telling title, "*How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions*", Forster and Sober (1994) apply Akaike's theorem, a theoretical result from statistics, to the problem of curve-fitting which involves a trade-off between goodness-of-fit and simplicity. From this, they

draw lessons for various philosophical debates, and concerning ad-hocness, they propose

> that a research programme is *degenerative* just in case loss in simplicity is not compensated by a sufficient gain in fit with data. Of course, the fit will always improve, but the improvement may not be enough to increase the estimated predictive value. (Forster and Sober, 1994, 17)

In a similar vein, Thagard (1988, 83) sees a negative relation between a theory's simplicity and the involvement of ad-hoc hypotheses. He even proposes a measure of explanatory power based on the consilience (number of facts explained given that the facts are equally important) and simplicity (depending on the number of explained facts and the number of "co-hypotheses"), which does not change if ad hoc elements (one additional "co-hypothesis" to explain one additional fact) are added (Thagard, 1988, 89–91).

In fact, we encountered similar elements in the present framework. Due to the liberal characterisation of (Account), elements may account for themselves. An ad hoc element is an element in a theory T that has no other purpose than accounting for itself. They exploit only the trivial inferential relation that holds between the element and itself. If this is the only inferential relation that pertains to an element of a belief system, its presence should be detrimental to the system's coherence, as the ad hoc element forms a subsystem that is not inferentially connected to the rest (BonJour, 1985, 97f).

Moreover, we can also see, that ad hoc elements trade-off simplicity and scope in a one-to-one manner: one additional element in the theory leads to one additional element in its scope. Granting more weight to syntactic simplicity than scope for their trade-off rectifies this problem. In such cases, ad hoc elements achieve account at the expense of unifying potential. Note that this addresses the issue of considering (Account) rather than (Residual Support'). The latter eradicates ad-hocness because an ad hoc element are accounted for only by themselves, and thus, they are not supported by the rest. In the present framework, we can deal with ad hocness if we supplement (Account) with (Syntactic Simplicity), (Scope) and an appropriate trade-off.

Concerning the relation between coherence and ad hocness, Schindler proposes following definition:

> A hypothesis H, when introduced to save a theory T from empirical refutation by data E, is ad hoc, iff (i) E is evidence for H and (ii) H appears *arbitrary* in that H coheres neither with theory T nor

> with background theories B – i.e., neither T nor B provides *good reason* to believe that H (possibly specifying a particular value of a variable) rather than non-H (or some value other than the one specified by H). (Schindler, 2018, 133)

Note the direction of Schindler's analysis: He uses coherence to define ad hocness but leaves coherence mostly unspecified adopting a pluralistic attitude towards the sources of theoretical reasons for believing a hypothesis (e.g., deduction, explanation or ruling out inconsistent alternatives) (Schindler, 2018, 134). I propose to take the opposite route. We can spell out coherence with theoretical virtues, and then show that ad hocness is detrimental to coherence due to impairing specific theoretical virtues or striking a bad balance between them.

5.3.3 Summary

Let us summarise and organise the coherential theoretical virtues according to the distinctions of pure and hybrid virtues, as well as necessary requirements and ideal desiderata (see Section 4.1). Consistency is a necessary requirement of coherence, internal consistency $\bot \notin Cn(T)$ is a pure theoretical virtue, external consistency is hybrid as it includes the inferential constraints in the background ($\bot \notin Cn(T \cup I)$). Account, simplicity and scope are desiderata that come in degrees. Simplicity is a purely theoretical virtue, scope and account are hybrid. Scope concerns the theory in relation to inferential constraints in the background, and account involves three components, theory, background and literals.

The three desiderata can pull in different directions to some extent. A simple theory may not be able to account for every literal, or a theory may achieve broad scope at the expense of simplicity. We can have broad scope and low account, but scope cannot be lower than account, every accounted for element is in the closure of a theory. This calls for trade-offs, a topic that we will address in the next chapter. For now, let us assume that we can strike a good balance, and that the degree of coherence of $B = L \cup I$, which now is required to be consistent, depends on a consistent theory $T \subseteq L$, and the degree to which T accounts for L, the degree to which T is simple, and the degree to which T is broadly scoped.[16]

[16]Note that this amounts to a gradual notion of *systemic* coherence, which pertains the belief base ($B = L \cup I$) as a whole. In contrast, there is also *relational* coherence, e.g., the degree to which T coheres *with* L. For work on the interdefinability of systemic and relational coherence, see Olsson (1999) and Hansson (2006).

My proposal sits quite well with other multicriteria approaches to coherence, and does not mark a complete departure from earlier works. I restrict myself to a comparison to the seminal account of BonJour (1985).[17]

BonJour (1985, 95–99) proposes the following criteria of coherence:

(i) A system of beliefs is coherent only if it is logically consistent.

(ii) A system of beliefs is coherent in proportion to its degree of probabilistic consistency.

(iii) The coherence of a system of beliefs is increased by the presence of inferential connections between its component beliefs and increased in proportion to the number and strength of such connections.

(iv) The coherence of a system of beliefs is diminished to the extent to which it is divided into subsystems of beliefs which are relatively unconnected to each other by inferential connections.

(v) The coherence of a system of beliefs is decreased in proportion to the presence of unexplained anomalies in the believed content of the system.

(i) makes consistency a necessary requirement for coherence. As (ii) involves probabilities, there is no counterpart in my propositional framework. The "presence of inferential connections" in (iii) is reflected in account. The discussion above revealed that we should not make any rule of deductive inference equally relevant to coherence (e.g., conjunction introduction or elimination). Account is an attempt to get hold of the more substantive deductive inference relations that obtain between the literals given the inferential background. BonJour motivates (iv) in the light of coherence requiring a belief system to form a "unified structure" (1985, 97). I presented unification to results from trading off simplicity and scope. (v) concerns explanatory relations, which may not be captured by deductive inferences alone. However, deductive anomalies ($\lambda \in L$ such that $Cn(T \cup I) \nvdash \lambda$) affect account negatively.

[17] A comparison with Thagard's approach of coherence as constraint satisfaction (2000) is more difficult. My proposal comes closest to his notion of "deductive coherence" (Thagard, 2000, 53).

5.4 Upshots for RE

Let us take a stance and see what upshots for RE arise from developing a virtue-based account of coherence in a simple deductive framework.

First, I think that we are now in a position to present additional motivation to include theoretical virtues in RE, as elaborate accounts of RE already do. Commonly, the motivation for the involvement of theoretical virtues stems from the idea of systematisation, and from drawing a parallel between theoretical deliberation in science and in RE. I suggest that we bridge the gap with coherence as an objective of RE that guides the configuration of virtues.

Looking back at philosophy of science, we can see that many prominent virtues occur in the present account as well. Moreover, there is significant overlap with virtues that are mentioned in the literature about RE. Consistency, account, simplicity, and scope figure as theoretical virtues that are generally relevant to RE. This quadruple of virtues (among many other virtues) figures in accounts of RE from the earliest descriptions in (Rawls, 1951) to the most recent application (Rechnitzer, 2022).

In addition, the present account proves to be helpful in disentangling complex virtues such as unification or non-ad-hocness by reducing them to a combination of account, simplicity and scope. I suggest, that striving to achieve unifying potential, and ultimately actual unification, is a way to spell out the idea of systematisation in RE.

Formalisation also counteracts the tendency to voice ideas about coherence vaguely. This is understandable on the highly general level of the philosophical debate about RE. However, it also threatens to result in paying lip service to coherence, or to result in an exercise of hand-waving. My attempt is not the first one. Interestingly, Goodman, who brought up the idea of RE (Goodman, 1983(1955)), also provides a formal approach to simplicity in the setting of predicate logic (e.g., Goodman, 1955). As the present framework is propositional in nature, a comparison with Goodman's much more elaborate treatment of simplicity is not feasible.

Finally, the virtue-based approach to coherence as an objective of RE can help to addresses worries of RE critics. Arras (2007, 58), for example, distinguishes two interpretations of coherence for moral justification in his critical discussion of RE. On a weak understanding, coherence is mere consistency among all elements, or more robustly, the notion is based on ideas from science. The robust notion of coherence draws its justificatory power from two features. First, it includes observation statements, which are supposed

to have firmer epistemological standing than considered moral judgements. Second, there is a "bootstrapping effect" of mutual support and "credibility transfer" from one subject matter of science to another. The moral domain, however, lacks a similarly strong relation of mutual support among judgements, principles and background theories. Thus, a dilemma arises: either coherence in RE as mere consistency is too weak to discriminate between live options in moral philosophy, which all pass this minimal test of coherence (2007, 59), or RE relies on a highly doubtful parallel to the epistemic standing of observations in science for a relation of mutual support.

The present virtue-based approach to coherence escapes Arras' dilemma. Spelling out one kind of support relation, deductive inference, and complementing it with theoretical virtues is clearly stronger than equating coherence with consistency. Moreover, my approach does not involve the arguably problematic reliance on the epistemic standing of commitments that has to be transferred to the entire system. Instead, the acceptability of a virtuous system is boosted as a result of theoretical virtues being conducive to the broadly epistemic objective of coherence. I suppose that this also has pragmatic consequences. In line with (Haslett, 1987), Arras claims that coherence considerations by themselves are of no help to decide which conflicting elements (judgements, principles) need to be adjusted, which results in "innumerable" different reflective equilibria (Arras, 2007, 59).[18] Arras sees little prospect to respond to this problem by appealing to relations such as "the best fit" or "the strongest mutual support". According to him, RE is a mix of disparate elements and various support relations for which it is not fleshed out how we could arrive to overall or maximal coherence rendering RE a useless guide (Arras, 2007, 60).

In this case, the upshot of having a simple but fully formalised account stems from coherence considerations that include an array of virtues, which provide more guidance than mere consistency or everything fitting together. Moreover, the formal framework facilitates the work towards aggregating and trading-off theoretical virtues towards an overall notion of virtuousness in the next chapter. Note that there will not be uniquely best systems according to a multicriteria approach such as the present account of virtue-based coherence, but it seems to me out of question that there are innumerably many equally virtuous alternatives.

[18] In the present framework, we have to understand "innumerable" as "many" outcomes because "innumerable" the the mathematically precise sense of *uncountably infinite* does not apply. The set of literals and belief bases are assumed to be finite, which allows for enumeration.

So, we have some leverage to address objections of conservativity and no-convergence. If RE involves the consideration of virtues that go beyond mere consistency and fit, namely simplicity and scope, and emerging from them, unification and non-ad-hocness, there is notable pressure to revise unorganised views about a subject matter that go far beyond conservative streamlining with minimal adjustments. Concerning convergence, theoretical virtues are a double-edged sword. On the one hand, we could hope that the guidance provided by virtues is convergence-conducive by restricting the set of viable alternatives during a process or in equilibrium. This might work if configurations of different agents are (roughly) the same. On the other side, disagreements about configurations provide an additional source for divergence. In this case, the upshot of virtues stems from being able to explain disagreement by backtracking differences in equilibria to differences in configurations.

Next, Bonevac (2004) objects to RE on the basis of BRT. Hence, the present approach may form the starting point to develop a positive outlook. However, this goes beyond the scope of the present project. For an early attempt in this direction, see (Freivogel, 2021).

To conclude this chapter, let us turn back to the list of epistemic desiderata of Kappel (2006). I am not going to contest Kappel's attack on the truth-conduciveness of theoretical virtues, as I attempt configure theoretical virtue towards the objective of coherence. Nonetheless, the present approach takes up many elements of Kappel's list of epistemic desiderata (Kappel, 2006, 132f)[19]. Consistency, simplicity and generality (scope) are obviously present, systematicity, too, if we interpret account to capture some aspects of explanatory relations. As the belief base is supposed to consist of beliefs that an agent holds *independently* of others, a weakly foundationalist factor is implemented in the present framework. This captures Kappels's desideratum of intuitive acceptability. Trade-offs are not a theoretical virtue but the framework of the virtue-based approach is ready to operationalise them. Finally, Kappel's last desideratum is in line with my view that the better a system instantiates theoretical virtues the more coherent it is. All in all, this illustrates that Kappel's suggested tasks from the beginning of the chapter to spell out coherence in terms of theoretical virtues and and "stating the desiderata more precisely and sorting out their interrelations if our aim were to provide a full defence of RE" (Kappel, 2006, 132) can be accomplished.

[19]The complete list and a discussion of its elements can be found on page 69.

This completes the first step of configuring theoretical virtues for RE. We set up the objective of coherence, selected and specified virtues in a simple propositional framework. In the next step, we need to address the aggregation theoretical virtues into an overall ordering of virtuousness that allows for trade-offs between virtues. I will take up this task in the next chapter.

Appendix

B.1 Proofs

Proposition 1. *Let $L = \{\lambda_1, \ldots, \lambda_n\}$ be a set of literals. $B = L \cup I$ satisfies (Residual Support') if and only if*

$$\lambda_k \vee \bigvee_{\substack{i=1 \\ i \neq k}}^{n} \neg\lambda_i \in Cn(I)$$

for all $k \in \{1, \ldots, n\}$.

Proof. We prove the two directions of the equivalence separately.

"⇒": Assume that $B = L \cup I$ satisfies (Residual Support'), that is,

$$L \setminus \{\lambda_k\} \cup I \vdash \lambda_k$$

for all $k \in \{1, \ldots, n\}$. By the deduction theorem of classical propositional logic, we have

$$I \vdash (\bigwedge_{\substack{i=1 \\ i \neq k}}^{n} \lambda_i) \rightarrow \lambda_k$$

for all $k \in \{1, \ldots, n\}$. This is equivalent to

$$I \vdash (\bigvee_{\substack{i=1 \\ i \neq k}}^{n} \neg\lambda_i) \vee \lambda_k$$

for all $k \in \{1, \ldots, n\}$, which is a notational variant of

$$\lambda_k \vee \bigvee_{\substack{i=1 \\ i \neq k}}^{n} \neg\lambda_i$$

for all $k \in \{1, \ldots, n\}$.

"$\Leftarrow$": Let $k \in \{1, \ldots, n\}$ and assume that

$$\lambda_k \vee \bigvee_{\substack{i=1 \\ i \neq k}}^{n} \neg\lambda_i \in Cn(I).$$

This implies

$$(\bigwedge_{\substack{i=1 \\ i \neq k}}^{n} \lambda_i) \rightarrow \lambda_k \in Cn(I).$$

As $L \smallsetminus \{\lambda_k\}$ provides the antecedents of this conditional, we have

$$L \smallsetminus \{\lambda_k\} \cup \{(\bigwedge_{\substack{i=1 \\ i \neq k}}^{n} \lambda_i) \rightarrow \lambda_k\} \vdash \lambda_k.$$

Consequently,

$$L \smallsetminus \{\lambda_k\} \cup I \vdash \lambda_k,$$

and because k is arbitrary, this holds for all $k \in \{1, \ldots, n\}$. This implies that $B = L \cup I$ satisfies (Residual Support'). □

Chapter 6

How (Not) to Aggregate and Trade Off Theoretical Virtues

In the previous chapter, I presented theoretical virtues in view of the objective of coherence, and spelled them out in a deductive framework: consistency, account, simplicity and scope. Now, having dealt with selection and specification, we turn to the second part of devising a configuration of theoretical virtues for RE: Aggregation and trade-offs. Some virtues are a categorical matter of yes or no, e.g., consistency or full account. Other virtues come in degrees, e.g., account, simplicity or scope, and they can pull in different directions. Hence, aggregation and trade-offs are directed at an evaluation of a theory's overall virtuousness. This forms the basis for theory choice, or theory adjustments in science as well as in RE.

The aim of this chapter is to provide a short piece of groundwork, which I conceive as indispensable. Kuhn (1977, 359) laments that no progress has been made towards aggregating and weighting criteria for theory choice in science, and I think, that this holds to this day of writing my dissertation (2023). The result of this chapter will not be a final solution for trading-off and aggregating theoretical virtues, but an illustration of obstacles, which might become relevant during configuring theoretical virtues for RE, and a how they might be overcome.

The groundwork comprises different tasks: First, it is important overcome vagueness and get a formally precise understanding of what it could mean to compare theories, and claim, for example, that one theory is simpler than other. Second, we need to recognise the limitations of what can be done reasonably with different kinds of orderings.

The chapter is structured as follows. In Section 6.1, I provide a hopefully gentle introduction to the topic of orders. In order to illustrate the aggregation of virtues, I present three strategies and apply them to a step in a recent and elaborate application of RE by Rechnitzer (2022). In Section 6.2, I present

the Arrowian impossibility results that have been transferred to theory evaluation in science by Okasha (2011) as an important limitation of what can be done reasonably with orderings on ordinal scales. I present possible escape routes. Next, I demonstrate that the three strategies fail Arrow's theorem in different respects. The observation that the strategies do not allow for trade-offs provides enough motivation to turn our attention to numerical measures and an additive aggregation function in the formal framework in Section 6.3.

6.1 A Primer on Ordering Theories with Virtues

Theories can instantiate gradual virtues to a higher or lesser extent. It seems quite natural to say that some theory is simpler than another, or that it has broader scope than its rivals. Consequently, such virtues provide an ordering of theories, each virtue on its own. Orderings can provide us with various degrees of information about its differently ranked elements. The most widely used categorisation in this respect distinguishes between nominal, ordinal, interval and ratio scales. A nominal scale is qualitative in nature and its label are used to identify under which category an element falls, for example "consistent" and "inconsistent". Nominal scales are not equipped with an inherent idea of an order among the categories. Before we move on to provide numerical measures for virtues on interval or ratio scales it is important to develop a more basic understanding of orders in terms of ordinal scales. If one is not ready to adopt the highly simplifying and idealising assumptions that go into formalisation and numerical measures, one can fall back on the groundwork.

Given a set of theories $\mathcal{T}$ and a name for a virtue v, a *non-strict preorder* $\preccurlyeq_v$ is a binary relation on $\mathcal{T}$. For two theories $S, T \in \mathcal{T}$,

$$S \preccurlyeq_v T$$

can be read as

$$T \text{ is as least as virtuous as } S \text{ with respect to virtue } v.$$

The *strict* part of $\preccurlyeq_v$ is denoted by $\prec_v$, and

$$S \prec_v T$$

stands for

T is (strictly) more virtuous than S with respect to virtue v.

Non-strict preorders satisfy the following conditions for all $R, S, T \in \mathcal{T}$:

(Reflexivity) $T \preccurlyeq_v T$

(Transitivity) If $R \preccurlyeq_v S$ and $S \preccurlyeq_v T$, then $R \preccurlyeq_v T$.

As it stands, there may be theories that are *incomparable* with each other, i.e., there may be $S, T \in \mathcal{T}$ such that neither $T \preccurlyeq_v S$ nor $S \preccurlyeq_v T$ holds. A preorder without incomparable pairings, is called *total*, and it satisfies, in addition to previous conditions, the following for all $S, T \in \mathcal{T}$:

(Totality) $T \preccurlyeq_v S$ or $S \preccurlyeq_v T$

Occasionally, total preorders are called "preference relations", and if we interpreted virtues as persons, we could read $S \prec_v T$ as "Virtue v (strictly) prefers T over S". A total preorder that satisfies the following condition is a *total order*:

(Antisymmetry) If $T \preccurlyeq_v S$ and $S \preccurlyeq_v T$, then $T = S$.

In the context of theory evaluation, (Antisymmetry) is an implausibly strong requirement, as it blocks ties between different theories with respect to some virtues.[1] If S and T are two equally virtuous theories with respect to v (denoted by $T \sim_v S$), then S is at least as virtuous as T ($T \preccurlyeq_v S$), and T is at least as virtuous as S ($S \preccurlyeq_v T$), but they may well be distinct ($S \neq T$).

One might also doubt, whether orderings according to theoretical virtues generally satisfy (Totality). Plausibly, an agent may face a situation, where they are not able to rank two rival theories on the basis of a theoretical virtue. In this case, the ordering o theories would need to be adjourned or to be made tentatively. However, for theory choice, the theories in a set of rivals $\mathcal{T}$ (e.g., the geocentric and the heliocentric theory in astronomy) typically have to address the same subject matter (e.g., celestial movement) and answer to the same evidence or explain the same phenomena (e.g., the retrograde motion). Typically, this narrows down the range of $\mathcal{T}$ to a few candidates, increasing the chances that the theories can be compared to each other.[2] Instead, failing

[1] In other contexts (Antisymmetry) is perfectly reasonable. The relation "$\leq$" ("... is smaller than or equal to ... ") on the integers satisfies (Antisymmetry), and indeed, is a total order.

[2] For a similar point, see (Morreau, 2014, 1256).

to establish a total preorder among a set of rival theories with respect to a theoretical virtue v, may also give us a reason to remove v from a selection of virtues for the time being, as v is not helpful to facilitate choice. In what follows, I will assume that gradual theoretical virtues establish total preorders on sets of theories.

Until now, each theoretical virtue is supposed to order a set of theories on its own. The tricky part is to aggregate those individual orders into a single ordering of "overall betterness" with respect to all virtues. Let us call a function, which, maps a list of n total preorders $\langle \preccurlyeq_{v_1}, \ldots, \preccurlyeq_{v_n} \rangle$, a so-called *profile*, to a single total preorder $\preccurlyeq$, an *aggregation rule*.[3] We may think of aggregation rules as a special kind of more general aggregation *strategies*, i.e., any kind of instructions to aggregate orderings. There are aggregation strategies that do not yield total preorders (see below for examples), and thus, do not belong to the narrower category of aggregation rules.

The task of aggregating virtue orderings is complicated by the fact that theoretical virtues can "pull in different directions", i.e., they may order theories differently. For example, take total preorders for simplicity $\preccurlyeq_{sim}$ and scope $\preccurlyeq_{sco}$ and two theories T and S. Suppose that T achieves broader scope than S at the expense of simplicity. Consequently, the following relations may obtain:

$$T \prec_{sim} S$$
$$S \prec_{sco} T$$

There is no straightforward solution to arrive at an overall order of virtuousness that dissolves the tension between the individual orderings according to simplicity and scope.

Before we discuss various approaches to solve the problem of aggregating individual orderings into an overall order of virtuousness, I introduce the notion of *Pareto efficiency* (also: Pareto optimality) that covers the clear-cut cases and leads to a restriction of viable candidate theories. Let v_i $(i = 1, \ldots, n)$ be theoretical virtues, and let $\mathcal{T}$ be a set of theories. We say that a theory $T \in \mathcal{T}$ *dominates* (or improves on) another theory $S \in \mathcal{T}$ if and only if $S \preccurlyeq_{v_i} T$ for all $i \in \{1, \ldots, n\}$, and there is some $j \in \{1, \ldots, n\}$ such that $S \prec_{v_j} T$. To put it less technically, a theory dominates an alternative if and only if it is at least as good as the alternative with respect to every virtue, and strictly better for

[3] They go by different names in the literature. Arrow (1951), for example, speaks of "social welfare functions", Okasha (2011) calls them "theory choice rules".

at least one virtue. A theory $T \in \mathcal{T}$ is *Pareto efficient* (Pareto optimal) if and only if it is not dominated by any other alternative in $\mathcal{T}$. In other words, an agent cannot switch from a Pareto efficient theory to an alternative without performing worse according to at least one virtue. Figure 6.1 illustrates the definitions for dominated and Pareto efficient theories.

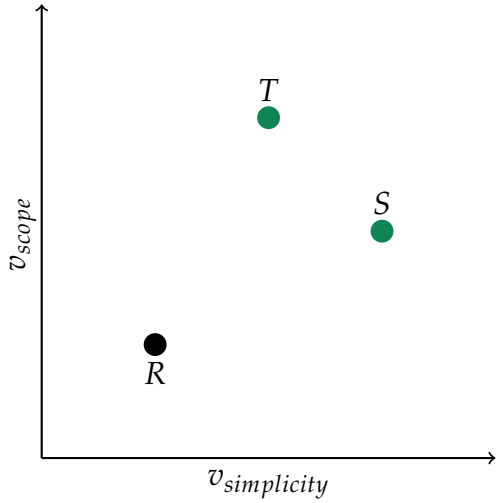


FIGURE 6.1: The visual placement of three theories (R, S, and T) corresponds to their positions in the orderings according to the virtues of simplicity (horizontal axis, increasing towards the right) and scope (vertical axis, increasing towards the top). R is dominated by both T and S. T and S do not dominate each other, and hence, are both Pareto efficient.

It is epistemically desirable to strive for a state that includes a Pareto efficient theory, otherwise, there is obvious room for improvement.[4] By changing from a Pareto inefficient to an efficient theory, one can improve with respect to at least one virtue without facing setbacks according to other virtues.

The notion of Pareto efficiency helps to screen off theories that are dominated by others, but it offers no resources to select among, reduce further, or bring more order to Pareto efficient theories.

Note that the idea of Pareto efficiency already figures in the characterisation of equilibrium states in elaborate account of RE. Recall the fourth condition from Section 2.2, which is inspired by Elgin (1996, 107), and taken up by Rechnitzer (2022, 35):

4. The position is at least as reasonable as any available alternative in light of the initial commitments.

[4] Aiming for a Pareto efficient *state* according to selected virtues is not to say that a *process* leading towards such a state cannot allow for intermediate stages with dominated theories or transient setbacks. For a similar point about positions in an RE process, see (Rechnitzer, 2022, 56).

Note that this condition also refers to the initial commitments, and hence comprises more than Pareto efficiency of a position's theory with respect to theoretical virtues. Nonetheless, "at least as reasonable as any available alternative" can be understood to signify Pareto efficiency in theory adjustment steps in a process of equilibration, for which we can assume that the commitments are provisionally held fixed.

Three Aggregation Strategies Let us turn to an illustrative selection of three aggregation strategies.[5] Later, we will see that they fall short of three different, and presumably, reasonable requirements for aggregation rules.

First, we can aggregate a profile of orderings by the *lexicographic* order. Here, we set out from a total order of the virtues itself, which we assume to be given, e.g.,

$$\text{account} > \text{simplicity} > \text{scope}.$$

Going from higher to lower ranked virtues, the overall order of two theories depends on the first virtue, for which the theories differ. The alphabetic ordering of names on a list is an example of a very common lexicographic order. If we have

$$T_1 \succ_{acc} T_2 \sim_{acc} T_3$$

and

$$T_2 \succ_{sim} T_1 \succ_{sim} T_3,$$

then $T_1 \succ_{lex} T_2$, because T_1 and T_2 differ with respect to account. T_2 and T_3 do not differ with respect to account, and hence, the next virtue, simplicity, is considered. Consequently, $T_2 \succ_{lex} T_3$.

The lexicographic order can aggregate profiles of orderings into total preorders. Hence, it can serve as an aggregation rule if the ordering of virtues is given. However, lexicographic orders effectively avoid trade-offs. No improvement in a lower ranked virtue can make up for a loss in a higher ranked virtue. In the present example, not even a huge gain in simplicity would allow for a minor setback in account.[6]

The best of the rather rare examples of theoretical virtues, which we may grant lexical priority over others, are necessary requirements in view of an epistemic-pragmatic objective. For example, consistency is frequently treated

[5] There are many more examples such as Condorcet's pairwise majority voting or Borda counting. For an overview of such strategies in social choice theory from a philosophical perspective, see List (2022).

[6] Because the virtuousness of a theory is not measured at this point, we have to understand "gain" and "setback" in terms of a theory's position in an order.

as a necessary requirement for coherence. However, such cases are rather rare and we still lack the ability to trade-off other virtues.

The second aggregation strategy concerns virtues that are presented with a *ceteris paribus* clause (for examples, see my discussion of unification in Section 5.3). Such clauses aim to emphasise the individual, positive contribution of every virtue or its ability to break ties. However, I lamented earlier that equipping formulations concerning theoretical virtues with "other things being equal" or its cognates grants very little practical help to establish an overall ordering. Now, we are in a position to spell this out formally.

Recall that two theories, which are equally virtuous (indifferent) with respect to a virtue v, are related by $\sim_v$. If a theory $T \in \mathcal{T}$ is overall more virtuous than $S \in \mathcal{T}$ if, other things being equal, T is more virtuous with respect to virtue v_j, then this amounts to $S \prec_{v_j} T$ and $S \sim_{v_i} T$ for all $i \in \{1, \ldots, n\}$, $i \neq j$. However, the second condition is very demanding, and probably rarely satisfied. As theoretical virtues can pull in opposite directions (e.g., simplicity vs. scope), it is implausible that we would arrive at an overall ordering of theories. Consequently, the aggregation instructions for formulations with *ceteris paribus* clauses amount to an aggregation strategy but not an aggregation rule.

Thirdly, there is, in my view, a very elegant refinement of Pareto efficiency by Das (1999), without involving more information than what is provided by the individual virtue orderings. Here, I transfer Das' results from objective functions f_i (typically real-valued) to virtues v_i ($i \in \{1, \ldots, n\}$) and their corresponding total preorders $\preccurlyeq_{vi}$. It goes as follows: Consider all k-element subsets of the n given virtues ($1 \leq k \leq n$). $T \in \mathcal{T}$ is *Pareto efficient of order* k if there is no $S \in \mathcal{T}$ that dominates T with respect to any of the k-element subsets of virtues.

Pareto efficiency of order n is ordinary Pareto efficiency, and a theory that is maximal with respect to every virtue would achieve Pareto efficiency of order 1. Consequently, Pareto efficiency of order k is an intermediate notion. It is stronger than ordinary Pareto efficiency, but weaker than maximal virtuousness with respect to all virtues. Geometrically, we might think of Pareto efficiency of order k to be Pareto efficiency in all k-dimensional subspaces of a n-dimensional virtue space. The lower k is, the better. Das (1999, 31) shows that every theory that is Pareto efficient of order k is also Pareto efficient of order j for every $n \geq j > k$. Considering refined Pareto efficiency results in a total preorder of theories, and thus forms an aggregation rule.

Take, for example, the following three virtue orderings for three theories:

$$T_1 \succ_{v_1} T_2 \succ_{v_1} T_3$$
$$T_2 \succ_{v_2} T_1 \succ_{v_2} T_3$$
$$T_3 \succ_{v_3} T_1 \succ_{v_3} T_2$$

All three theories are Pareto efficient (ordinarily, or equivalently, of order 3). T_3 is dominated by T_1 and T_2 for the two-element subset of virtues consisting of v_1 and v_2. Hence, T_3 is not Pareto efficient of order 2. Neither is T_2, as it is dominated by T_1 for v_1 and v_3. Finally, T_1 is not dominated by any other theory for any two-element subset of virtues, and hence Pareto efficient of order 2. This gives us a reason to prefer T_1 over T_2 and T_3 even though all are Pareto efficient in the unrefined sense.

An Illustrative Example Rechnitzer (2022) provides an extensive and by far the most detailed case study of an application of RE to the justification of a precautionary principle. Her elaborate account of RE falls in line with ideas voiced by Elgin (1996, 2017) and Brun (2020). In particular, steps in the RE process that involve the choice or the adjustment of theories are guided by theoretical virtues. She includes the virtues of determinacy, practicability, simplicity and scope, which are configured towards the pragmatic-epistemic goal of "justifying an action-guiding moral principle that is applicable to the subject matter of precaution and precautionary decision-making." (Rechnitzer, 2022, 101). She requires that theoretical virtues are comparable on at least ordinal scales (Rechnitzer, 2022, 54f), which is to say that they give rise to total preorders.

Theory adjustment step A_4 (Rechnitzer, 2022, 169–179) is of particular interest to us.[7] At this stage of the process, one has to choose among five candidate theories ("systems"): $RCPP, UUP, MPP, TPA$ and $P3$. Details need not concern us here except that they are ordered according to theoretical virtues

[7] I select this step for its wide field of candidate theories. I do not claim that the following considerations yield the same results in other steps as well. The present examples serves as an illustration for aggregation rules and not as a full reconstruction of Rechnitzer's RE process.

as follows (Rechnitzer, 2022, 176):

$$\begin{aligned}
\text{determinacy:} &\quad UUP \sim_{det} MPP \succ_{det} TPA \succ_{det} RCPP \succ_{det} P3 \\
\text{practicability:} &\quad MPP \succ_{pra} UUP \succ_{pra} TPA \succ_{pra} RCPP \succ_{pra} P3 \\
\text{scope:} &\quad TPA \sim_{sco} P3 \sim_{sco} MPP \succ_{sco} UUP \succ_{sco} RCPP \\
\text{simplicity:} &\quad UUP \succ_{sim} RCCP \sim_{sim} MPP \succ_{sim} TPA \succ_{sim} P3
\end{aligned}$$

and with respect to account: (Rechnitzer, 2022, 179)

$$\text{account:} \quad TPA \succ_{acc} P3 \succ_{acc} MPP \succ_{acc} RCPP \succ_{acc} UUP$$

Given the scope and complexity of Rechnitzer's application, as well as her view that the weighing of virtues can change during the process (2022, 57), she does not rely on a fixed mechanical aggregation rule and proceeds to aggregate and trade-off theoretical virtues on a case-by-case basis. She follows a strategy to first derive partial orderings from pairwise comparison of theory candidates with respect to the four theoretical virtues. Subsequently, she arrives at an overall comparison which also includes account.[8] Theory adjustment step A_4 results in the adoption of MPP as current system (Rechnitzer, 2022, 179)

How do the three examples of aggregation strategies perform if we apply them to Rechnitzer's profile? The aggregation instruction based on *ceteris paribus* clauses is completely useless for the present profile. Equally well performing theories (indifference relation $\sim$) occur far too sparsely to be exploited effectively. In order to get the lexicographic order off ground, we first need an among the virtues themselves. For the sake of illustration, assume that the order among virtues is

$$\text{determinacy} > \text{practicability} > \text{scope} > \text{simplicity},$$

which is roughly in line with what Rechnitzer (2022, 119) seems to have in mind at the outset of the process.[9] The first virtue, determinacy is indifferent about UUP and MPP, and thus, the second virtue, practicability, is considered. It yields MPP as the best theory according to the lexicographic order

$$MPP \succ_{lex} UUP \succ_{lex} TPA \succ_{lex} RCPP \succ_{lex} P3.$$

[8] Note that Rechnitzer does not treat account as a theoretical virtue.
[9] In addition, interchanging determinacy and practicability will yield the same result.

Finally, we can turn to refined Pareto efficiency. But first, there is a problem with Rechnitzer's definition of a Pareto optimal system, i.e., "a candidate [...] that is at least as good as all other alternatives with respect to all criteria, and better in at least one" (Rechnitzer, 2022, 56). A comparison with the definition from above, which I take to be the standard, reveals that Rechnitzer's definition is much more demanding. In fact, only one theory that is maximal with respect to every virtue ordering, could qualify as Pareto optimum according to Rechnitzer's definition.

It does not come as a surprise that Rechnitzer notes that there are no Pareto optima among the five candidate systems with respect to the four virtues. However, if we adopt the less demanding standard definition from earlier, there are Pareto optima: *MPP* and *UUP*.[10] *RCPP*, *TPA* and *P*3 are dominated by *MPP*. If we consult account afterwards to compare *MPP* and *UUP*, the former comes out on top. This is in line with Rechnitzer's result of adjustment step A_4 even though the result is reached in a different way.

We can arrive at the same conclusion if we consider all theoretical virtues including account. In this case, there are more Pareto optima: *MPP*, *UUP* and *TPA*. At this point, the order of Pareto efficiency comes in handy. *MPP* and *UUP* are Pareto efficient of order 4, and *TPA* is dominated by *MPP* for the four-element subset of virtues consisting of determinacy, practicability, scope and simplicity. Next, *MPP* is also Pareto efficient of order 3, and *UUP* is dominated by *MPP* for the three-element subset with determinacy, scope, and account.[11] Consequently, *MPP* should be preferred from the viewpoint of refined Pareto efficiency with respect to all virtues including account. This shows that we can arrive at *MPP* with an aggregation rule that does not rely on establishing partial orderings that leave room for judgement whether a theory's virtuousness can make up for low account values or not.

In view of objections claiming that RE is uninstructive because it "leaves you to muck about" how to resolve conflicts (Foley, 1993, 128), I think that it is extremely important to demonstrate that trading off theoretical virtues in RE can be spelled out in a formally rigorous manner. This is not to say that RE always has to be that formally rigorous in applications. However for the sought after proof of concept in this project, formal rigour is required.

[10] As the bookkeeping of theories, virtues and their relations, is a cumbersome and error-prone task, even for small examples, I devised a simple tool that is able to search for Pareto optima (of order *k*). It is available at https://github.com/free-flux/virtuously-circular/tree/main/chapter-6, and it produced the results reported here.

[11] In fact, *MPP* is "almost" Pareto efficient of order 2. It is dominated by alternatives for two out of ten two-element subset of virtues. Das (1999, 32) calls this a Pareto optimum of order 2 with degree 8.

Even though that the notion of refined Pareto efficiency plays out very well in the illustrative treatment of Rechnitzer's adjustment step A_4, there remains a serious shortcoming. It does not allow to weigh virtues and their preorders differently. Every virtue receives equal amount of attention in the assessment towards Pareto efficiency of order k. Das (1999, 32) observes that theories with lower orders of efficiency tend to have less extreme positions in individual virtue orderings. This may result in some trade-offs (e.g., *MPP* striking the best balance between account and the rest of theoretical virtues), but we still lack full control of trade-offs. The same issue applies to other aggregation rules as well.

To make things worse, this practical issue of lacking means to weigh virtues is joined by more theoretical problems in the next section.

6.2 Lessons from Arrow

Arrow (1951) proves a famous impossibility result in social choice theory which establishes that ordinal-ranked preferences cannot be aggregated by a so called "social choice function" into a single preference order that respects reasonable requirements. Briefly, the requirements go as follows (cf. Okasha, 2011, 89f):

(U) Unrestricted domain: The social choice function accepts all profiles, i.e., all lists of individual preferences, as inputs.

(P) Weak Pareto: If all individuals strictly prefer x to y, the aggregated social order should also prefer x to y.

(I) Independence of irrelevant alternatives: The aggregated social order of x and y only depends on the preferences of individuals between x and y and not on individual preferences concerning other alternatives.

(N) Non-Dictatorship: There is no individual such that if the individual strictly prefers x to y, then so does the aggregated social preference order.

Okasha (2011) transfers Arrow's theorem to theory choice by identifying the preferences of social agents with orderings according to theoretical virtues. The choices concern no longer social alternatives but scientific theories. In other words, each theoretical virtue acts as an agent with a preference ordering of alternative theories, which have to be aggregated into an overall

ranking of theories. If the requirements stated in Arrow's theorem plausibly hold for theory choice, then a theory choice function that respects all requirements is impossible.

Moreover, Okasha (2011) relates this result to Kuhn's issue of trade-offs. Kuhn (1977) suggest that there is no neutral, shared algorithm, because there is no subject-independent way to weigh and to trade off theoretical virtues. Kuhn does not exclude that there may be many algorithms of theory choice. In contrast, Okasha's application is an escalation, since Arrow's theorem implies that there is no reasonable algorithm at all.

It lies in the nature of such impossibility results that they spark a lively debate about possible "escape routes". This holds for (Arrow, 1951) as well as for (Okasha, 2011). For example, one could reject at least one of the four requirements as implausible. For theory choice, Morreau (2014, 2015) argues against Okasha that some theoretical virtues are "rigid", i.e., they order the alternatives in a single way, leading to the rejection of the requirement of an unrestricted domain (U). See Okasha (2015) for a reply.

As a corollary, the aggregation strategies discussed above, also have to violate at least one of Arrow's requirements. If we wanted to construe an aggregation rule and not just a strategy from virtues with ceteris paribus clauses, we could restrict the domain to suitable profiles.[12] This restriction of the domain of profiles violates (U). Next, lexicographic orderings make the highest ranked virtue a dictator in violation of (D). Finally, refined Pareto efficiency violates requirement (I). Put more formally as above, (I) states the following for any profiles $\langle \preccurlyeq_{v_1}, \ldots, \preccurlyeq_{v_n} \rangle$ and $\langle \preccurlyeq'_{v_1}, \ldots, \preccurlyeq'_{v_n} \rangle$, and any theories T_1 and T_2: If $\langle \preccurlyeq_{v_1}, \ldots, \preccurlyeq_{v_n} \rangle$ and $\langle \preccurlyeq'_{v_1}, \ldots, \preccurlyeq'_{v_n} \rangle$ are identical when restricted to $\{T_1, T_2\}$, then the overall rankings that arise from those profiles, $\preccurlyeq$ and $\preccurlyeq'$, are identical when restricted to $\{T_1, T_2\}$.

Consider the following situation (which is analogous to an example of Okasha, 2011, 93) for two theories, T_1 and T_2: T_1 is simpler, and has broader scope than T_2. In turn, T_2 is more accurate than T_1. Next, assume that these relations are embedded in two different profiles that involve a third theory T_3 depicted in Table 6.1.

According to profile A, T_1 is Pareto efficient of order 2, T_2 is Pareto efficient of order 3, and T_3 is dominated by T_2 (and hence, not Pareto efficient of any order). In this case T_1 is overall preferred over T_2. For profile B, things

[12] A suitable profile $\langle \preccurlyeq_{v_1}, \ldots, \preccurlyeq_{v_n} \rangle$, would prefer a theory T over all alternatives with respect to a single virtue v_i while all other virtues are indifferent towards all pairings ($T_1 \sim_{v_j} T_2$ for all $T_1, T_2 \in \mathcal{T}$ and $i \neq j$).

	profile A					profile B				
simplicity	T_1	$\succ_{sim}$	T_3	$\sim_{sim}$	T_2	T_1	$\sim_{sim}$	T_3	$\succ_{sim}$	T_2
accuracy	T_2	$\succ_{acc}$	T_3	$\succ_{acc}$	T_1	T_2	$\succ_{acc}$	T_3	$\succ_{acc}$	T_1
scope	T_1	$\succ_{sco}$	T_3	$\sim_{sco}$	T_2	T_1	$\sim_{sco}$	T_3	$\succ_{sco}$	T_2

TABLE 6.1: Two profiles of virtue orderings exhibiting the same relations between T_1 and T_2. T_1 is simpler, and has broader scope than T_2. T_2 is more accurate than T_1.

stand differently: Both T_2 and T_3 are Pareto efficient of order 3, but T_1 is dominated by T_3 (and hence, not Pareto efficient of any order). Thus, for profile B, T_2 is overall preferred over T_1. This violates (I) because both profiles exhibit the same virtue orderings between T_1 and T_2.

This is not that surprising, as Pareto efficiency (with or without orders) is defined to take *all alternatives* into account. So, we have to expect that differences in profiles that concern "irrelevant" alternatives (e.g., relations to T_3 in the example above) affect the overall ordering of theories according to Pareto efficiency. The question then is, whether we should see violation of (I) as a fatal flaw. After all, (I) is considered to be requirement for a reasonable aggregation rule. Unfortunately, the discussion goes far beyond the scope of this project. For a recent overview of criticism and defences of (I), see (Patty and Penn, 2019). Note that (I) is an *interprofile* requirement as it relates multiple profiles. In view of single profiles, e.g., Rechnitzer's profile in step A_4, (I) is irrelevant because there are no other profiles under consideration. However, Morreau (2014, 1265) states analogues for intraprofile requirements that yield similar impossibility results for single profile approaches.

6.3 Measuring Theoretical Virtues

There is only so much we can do to aggregate orderings if we have merely ordinal scales at our disposal for comparisons. Another escape route from Arrow's theorem arises from the assumption that the preferences are given as ordinal rankings that provide no information about the intensity of differences. This "informational basis" of theory choice can be enriched by considering more than ordinal rankings, e.g., measuring theoretical virtues on interval or ratio scales and allow for inter-comparability of theoretical virtues. In the field of welfare economics, such approaches have been studied extensively by Sen (e.g., 1970), and Okasha (2011) also transfers these results to the context of theory choice.

In the case of the formal framework of virtue-based coherence from the previous chapter, we are in the comfortable situation that we can gather more than ordinal information about the virtuousness of theories with respect to virtues that allow for degrees. Recall the formulations pertaining to account, simplicity and scope of the previous chapter. Let $B = L \cup I$ be a belief base consisting of a set of literal beliefs L and inferential beliefs I. L is assumed to be minimally consistent, i.e., it does not contain blatantly contradicting literals p and $\neg p$, or equivalently, $\bot \notin Cn(L)$. Recall, that $\mathfrak{L}$ denotes the set of literals, i.e., atoms and their negations that are relevant to the subject matter of inquiry. $n = \frac{1}{2}|\mathfrak{L}|$ is the number of relevant atoms and it serves to normalise the measures. Let $T \subseteq L$ be a non-empty theory, in particular we require that it is consistent given the inferential background I, i.e, $\bot \notin Cn(T \cup I)$. We denote the deductive closure of a theory (given background I) restricted to literals by $\overline{T}$, i.e., $\overline{T} = Cn(T \cup I) \cap \mathfrak{L}$.

(Account) The degree to which T accounts for L is proportional to the number of $\lambda \in L$, such that $T \cup I \vdash \lambda$.

(Syntactic Simplicity) The degree of syntactic simplicity of T is inversely proportional to $|T|$.

(Scope) The scope of T is proportional to $|\overline{T}|$.

The frameworks allows for simple counting as a first step. In particular, we can count the number of literals which a theory accounts for ($|\overline{T} \cap L|$), the elements in a theory ($|T|$) for syntactic complexity, and the literals in the closure of the theory ($|\overline{T}|$) for scope. There is a "natural" zero point of counting sentences: the empty set containing no elements at all. Consequently, we can devise measures on ratio scales, so that it is meaningful to state that "T_1 is twice as syntactically complex as T_2" or that "T_1 accounts for half as many commitments as T_2".

$$account(T, L) = \frac{|\overline{T} \cap L|}{|L|}$$

$$simplicity(T) = \frac{n - |T|}{n - 1}$$

$$scope(T) = \frac{|\overline{T}| - 1}{n - 1}$$

The functions are designed to yield values between 0 (worst) to 1 (best), which is achieved by a correction term (-1) for simplicity and scope.[13] For example, $scope(T) = 1$ if and only if the closure of T covers every atomic sentence ($|\overline{T}| = n$). Conversely, $scope(T) = 0$ if T contains a single element and does not exploit but the trivial inferential relationship ($|\overline{T}| = 1$).

Account is normalised by $|L|$ and not by n, because the latter would penalise narrowly scoped theories. This should be covered by scope in order to separate the related virtues of scope and account as good as possible. The measure for account remains very crude. It does not keep track of elements that go beyond L ($\lambda \in \overline{T}$ but $\lambda \notin L$), and it does not distinguish between unaccounted elements ($\lambda \in L$ but $\lambda \notin \overline{T}$) and contradicting elements ($\lambda \in \overline{T}$ but $\neg\lambda \in L$). Arguably, contradicting elements should have more negative impact on account than absent elements. These function are not the only way to operationalise virtues on a ratio scale. We might, for example, opt for non-linear function and such decisions would need to be explored in great detail if we aimed for more than a proof of concept.

Operationalising degrees of virtue instantiation with numerical measures induces total preorders naturally, as theories can be ordered according to their position on the number line for the value of their virtue measure $v()$. In this case , we have $T \preccurlyeq_v S$ if $v(T) \leq v(S)$, where $\leq$ is the usual relation "... is less than or equal to ..." for real-valued numbers.

A straightforward solution for aggregating the individual measure is to take their sum:

$$account(C, T) + simplicity(T) + scope(T)$$

The legitimacy of this move rests on the assumption that the individual measures are *commensurable*. For example, number of sold items of a product and the number of positive reviews may both be indicative of a product's popularity, but as numerical measures, they are incommensurable, Their sum is meaningless because they apply different units of measurement. Note that the counting of elements (sentences in sets) vindicates the assumption of commensurability in the present framework to some extent.

[13] Note that these functions are not "measures" in the mathematically strict sense. For this, they would need to be non-negative, assign 0 to the empty theory and the measure of the union of a countable number of pairwise disjoint theories would need to equal the sum of their measures.

Next, the issue of trade-offs can be addressed by introducing relative weights that are applied to the measure before summation:

$$w_1 \cdot account(C,T) + w_2 \cdot simplicity(T) + w_3 \cdot scope(T)$$

Imposing the boundary condition $w_1 + w_2 + w_3 = 1$ ensures that the resulting value falls in the range between 0 and 1.

Let us consider the toy example from Section 5.2, i.e.,

$$B = L \cup I = \{p, q, \neg r, \} \cup \{\neg p \vee \neg q \vee \neg r\},$$

to compare $T_1 = \{p, q\}$ and $T_2 = \{\neg r\}$ in Table 6.2. Note that we have $n = 3$ in the present example. Moreover, $\overline{T}_1 = \{p, q, \neg r\}$, $\overline{T}_2 = \{\neg r\}$. T_1 and T_2, are Pareto efficient, as neither one dominates the other. For different weightings, either one can come out on top. For $w_1 = w_2 = w_3 = \frac{1}{3}$, the sum of weighted measures is $\frac{5}{6}$ for T_1, and $\frac{4}{9}$ for T_2. In this case, T_1 is preferable over T_2. In contrast, for $w_1 = w_3 = \frac{1}{10}$ and $w_2 = \frac{8}{10}$, T_1 yields $\frac{3}{5}$ as the sum of weighted measures, while T_2 achieves a better value of $\frac{5}{6}$.

	T_1	T_2
$account(C,T)$	1	$\frac{1}{3}$
$simplicity(T)$	$\frac{1}{2}$	1
$scope(T)$	1	0

TABLE 6.2: Numerical values for account, simplicity and scope functions in the formal framework for a toy example.

Note that the aggregation of weighted measures operationalises the "overall virtuousness" of a belief base, but it is not a measure for coherence. Necessary requirements for coherence, such as consistency, and probably full and exclusive account, are not guaranteed, even if a belief base is highly overall virtuous with respect to gradual virtues. This is an asset as it gives opportunity to restrict weights to those which are coherence-conducive. If there are weights that prove to rank belief bases highly, which satisfy the necessary requirements, then these weights are preferable over those which do not. Nevertheless, there is no reason to think that these weights are uniquely determined. They depend on other epistemic-pragmatic objectives, and they may be subject to change during inquiry.

For a recent application of an additive aggregation function with weights for the evaluation of logical theories, see (Priest, 2019), and (Priest, 2016) for

its general discussion. Even more noteworthy is (Priest, 2001), where he outlines a formal framework of BRT that incorporates weighted aggregation to yield (partial) orderings among options for revised sets of beliefs.

Note that the present aggregation of weighted functions is just one, and a highly idealised approach to solve the issue of aggregating and trading-off multiple theoretical virtues. For a presentation of methods of multicriteria decision analysis at book length, see (Ishizaka and Nemery, 2013), for example. It goes beyond the scope of the present project to examine, which of these approaches could be fruitful in an less idealised RE setting.

Now, it is time for a change of scene: We will leave behind the BRT-framework for virtue-based coherence and move to a full-fledged model of RE in the formal framework of the *Theory of Dialectical Structures* (TDS) in the next chapter. This is mainly serves to short-circuit shortcomings of the present approach. For example, the present framework does not identify sets of initial or current commitments, and it does not spell out adjustment operations. I am confident that these problems could be overcome by further work. However, it will turn out that there is a close correspondence between the two frameworks, anyway. This allows to transfer insights about virtue-based coherence to the new framework. Conversely, the full-fledged model of RE in TDS could guide the further development of a BRT model of RE.

Chapter 7

A Formal Model of Reflective Equilibrium

An important tool and the driving force behind the rest of this project is the formal model of RE from Beisbart, Betz, and Brun (2021).[1] Their groundwork also serves as a proof of concept because they show that important RE elements can be specified in a consistent manner that renders RE operational. In particular, the formal model arrives at an operational configuration of theoretical virtues. Moreover their formal model is extremely fruitful as opens up many lines of research at the nexus of RE and formal modelling. It will provide the basis to explore whether we can address objections to RE with theoretical virtues by means of computer simulations.

The aim of the this chapter is to take the time to present the formal model in some detail. On the one hand, this task will recapitulate the work of Beisbart, Betz, and Brun (2021) to some extent, but it allows us to introduce important terminology, which we will use frequently throughout the project in Section 7.1, and highlight the involvement of theoretical virtues in the model in Section 7.2.

[1] At this point, there is only Thagard's formal model of coherence as constraint satisfaction, which has been discussed under the label of "reflective equilibrium" (Thagard, 2000, but see also Yilmaz, Franco-Watkins, and Kroecker, 2017). However, it bears a faint resemblance to informal accounts of RE. It does not distinguish between commitments and theory, and the process of equilibration is a connectionist algorithm of activation updating on a network of units connected by excitatory and inhibitory links rather than mutual adjustments.

7.1 The Formal Model

7.1.1 The Framework: Theory of Dialectical Structures

The model's framework of formalisation is called *Theory of Dialectical Structures* (TDS) (Betz, 2010, 2012). TDS is an approach to argumentation theory with applications for debate reconstruction, analysis, and evaluation of real controversies. A *dialectical structure* is a pair $\tau = (\mathcal{S}, \mathcal{A})$, where $\mathcal{S}$ is a sentence pool, and $\mathcal{A}$ is a set of all arguments.[2] The sentence pool $\mathcal{S}$ is the source of atomic elements representing natural-language sentences relevant to a subject matter. Sentences are represented by propositional variables (Roman lower-case letters with indices) without their inner logical or semantical relations. It is assumed that $\mathcal{S}$ is fixed and that it is closed under negation, i.e., $s \in \mathcal{S}$ implies $\neg s \in \mathcal{S}$. In addition, $\neg\neg s$ and s are identified by stipulation. $n = \frac{1}{2} \cdot |\mathcal{S}|$ is the size of the (unnegated) sentence pool and we frequently use it for normalisation.

Next, we assume that deductively valid arguments represent the inferential relations among sentences from $\mathcal{S}$.[3] An argument $\mathfrak{a} = (P_\mathfrak{a}, c_\mathfrak{a}) \in \mathcal{A}$ consists of a set of premises $P_\mathfrak{a} \subseteq \mathcal{S}$ and a conclusion $c_\mathfrak{a} \in \mathcal{S}$.

Given a dialectical structure $\tau = (\mathcal{S}, \mathcal{A})$, an agent can adopt a *dialectical position* P consisting of accepted sentences, e.g., $\{s_1, \neg s_2, s_3\}$.[4] A position P is called *complete* if and only if s or $\neg s$ is an element of P for every sentence s from the unnegated half of the sentence pool. If an agent maintains a complete position, there are no sentences from the sentence pool on which they remain silent. Otherwise, P is a *partial* position. Using the handy set notation, we say that a position Q *extends* another position P if and only if $P \subseteq Q$.

A position P is *minimally consistent* if and only if it does not contain a flat contradiction, i.e., a sentence s and its negation $\neg s$. A minimally consistent and complete position is *dialectically consistent* on a dialectical structure $\tau =$

[2] Beisbart, Betz, and Brun (2021) rely on a slightly simplified version of TDS. The full-fledged framework of Betz (2010, 2012) defines dialectical structure as a triple of arguments (from a sentence pool), an attack, and a support relation. In the present project, sentence pools and arguments suffice to model RE.

[3] More exactly, an argument represents multiple inferential relations due to contraposition. If $(\{s_1, s_2\}, c)$ is a deductively valid argument, then so are $(\{\neg c, s_1\}, \neg s_2)$ and $(\{\neg c, s_2\}, \neg s_1)$.

[4] Again, this is a handy simplification in comparison to (Betz, 2010, 2012). There, a position is a truth value assignment $\mathcal{P} : S \to \{\mathbf{t}, \mathbf{f}\}$, where $S \subseteq \mathcal{S}$. For minimal consistency, it is required that contradictory sentences are assigned opposite truth values. In the present context, this requirement is stipulated for simplicity's sake, which allows to declare positions on half of the sentence pool. Thus, we can identify a position P with the set of sentences which are assigned the truth value $\mathbf{t}$. Consequently, a position's set notation P and the truth value assignment $\mathcal{P}$ are related by $P = \mathcal{P}^{-1}(\mathbf{t})$, i.e., P is the preimage of $\mathbf{t}$ under $\mathcal{P}$.

$(\mathcal{S}, \mathcal{A})$ if and only if for every argument $\mathfrak{a} = (P_\mathfrak{a}, c_\mathfrak{a}) \in \mathcal{A}$: if $P_\mathfrak{a} \subseteq P$, then $c_\mathfrak{a} \in P$. Less formally, an agent in a complete and dialectically consistent position accepts all conclusions of arguments, for which they accept the premises. A partial position Q is dialectically consistent if it is minimally consistent and if there is a complete and dialectically consistent position P that extends Q.

The *dialectical closure* $\overline{P}$ of a dialectically consistent position P is the intersection of all complete and dialectically consistent extensions of P (denoted by $E(P)$):

$$\overline{P} = \bigcap_{Q \in E(P)} Q$$

The dialectical closure closes a position under all inferential relations arising from arguments, e.g., contrapositions or transitive "chains". Note that the dialectical closure of a position is again a position.

7.1.2 Formalising RE Components

Subject Matter RE aims at the justification of views, that revolve around a particular subject matter (e.g., inductive inference, a just setup of political and social institutions, or trolley cases). Moreover, the delineation of a particular subject matter is reasonable if we aim for coherence. Vastly diverse and unrelated views about various subject matters are hardly brought to cohere with each other, if coherence goes beyond mere consistency.

Thus, a fixed pool of sentences, which are relevant to a topic, is used to represent a subject matter. Typically, the sentences are inferentially related. Arguments that are built up from the sentence pool represent these relations. Consequently, the subject matter and the inferential relations formalised by a sentence pool $\mathcal{S}$ and arguments $\mathcal{A}$, respectively, give rise to a dialectical structure $\tau = (\mathcal{S}, \mathcal{A})$.

The dialectical structure, which serves as an example in (Beisbart, Betz, and Brun, 2021) is depicted in Figure 7.1. It consists of a sentence pool which comprises seven atomic sentences, $s_1, \ldots, s_7$. I refer to them by their integer indexes $(1, \ldots, 7)$ for the sake of simplicity. Negative numbers denote negated sentences. The dialectical structure contains the following single-premise arguments:

$$\mathcal{A} = \{(1,3), (1,4), (1,5), (1,-6), (2,-4), (2,5), (2,6), (2,7)\}$$

I will refer to it as the "standard example", and continue to use it. Its simple structure, which is depicted in Figure 7.1, allows to test intuitions, and do

calculations by hand.

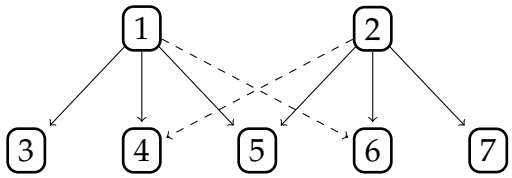


FIGURE 7.1: The dialectical structure of the standard example in (Beisbart, Betz, and Brun, 2021). Solid-line arrows indicate that the origin (premise) implies the target (conclusion), e.g., $(1, 3)$. Dashed-line arrows signify that the premise implies the negated sentence, e.g., $(2, -4)$.

Set of Commitments An agent expresses their view about a particular subject matter by accepting some sentences (or their negation which corresponds to rejection) that are relevant to the subject matter. An agent may not be committed to every sentence from a subject matter, and hence remain silent on some elements.

The formal model represents the set of commitments of an agent by a minimally consistent position C that is either complete or partial. Minimal consistency is a basic requirement of rationality, demanding that an agent does not explicitly accept a sentence as well as its negation. Note that this requirement does not rule out the dialectical inconsistency of commitments. An agent with minimally consistent commitments may still engage in contradictions given the inferential relations from the arguments of a dialectical structure.

Consider the following positions in the dialectical structure of the standard example:

$$C = \{3, 4, 5\}$$
$$C' = \{2, 3, 4, 5\}$$

Both positions are partial because they remain silent on 1 or 7, for example. C is dialectically consistent, but C' is not due to contradiction that arises from the argument $(2, -4)$.

Commitments and theories may change during an RE process, and thus, we use indices $i = 0, 1, 2, \ldots$ to discriminate between different stages. The *initial commitments* C_0 represent the starting point of an agent. Note that the model does not represent the idea of emerging commitments (Rechnitzer,

2022), i.e., underived commitments that arise during an RE process due to new information.

Theories The next important ingredient to RE are theories that serve to account for and systematise the set of commitments. Every sentence from a subject matter is admissible to act as an element of a theory. Theories are represented in the formal model by partial or complete positions that are dialectically consistent.

$\overline{T}$ is called the *content* or the *closure* of a theory T. Note that this motivates requirement of dialectical consistency for theories because the dialectical closure is not defined for inconsistent positions.

Consider the following positions in the standard example:

$$T = \{1, 7\}$$
$$T' = \{1, 2\}$$

T is dialectically consistent, and its closure is $\overline{T} = \{1, -2, 3, 4, 5, -6, 7\}$. In contrast, T' is not dialectically consistent given the arguments $(1, 4)$ and $(2, -4)$, and hence does not qualify as a theory.

Similar to commitments, theories are subject to change, and hence, indexed. We assume that the initial theory T_0 is empty.[5]

Epistemic States An *epistemic state* (C, T) is an ordered pair consisting of dialectical positions C and T that are required to be minimally and dialectically consistent, respectively. An epistemic state represents what has been called "position" in the informal account of RE in Chapter 2. I hope that the difference between the informal use of an agent's "position" for the pair a set of commitments and a theory, and the formal use of "dialectical position" and "epistemic state" becomes clear from the context.

Desiderata Three "forces" from the elaborate account of RE presented in Chapter 2 are represented in the formal model: fit between commitments and theory, respecting the input commitments, doing justice to theoretical virtues (the green arrows in Figure 2.1 of Chapter 2). They are modelled by three corresponding desiderata, namely *account*, *faithfulness*, and *systematicity*. The use of desiderata stresses the idea that they can be satisfied to a greater or

[5] It is a simplifying assumptions that an agent has not attempted or succeeded to systematise her commitments at the outset of RE. However, nothing in the formal model would prevent us to initiate RE processes from a non-empty theory.

lesser degree in contrast to necessary requirements that are a matter of all-or-nothing, e.g., minimal consistency for commitments. Consequently, we need to render the desiderata graded and the most straightforward way to do so is to introduce measures on a ratio scale for them.

Account concerns the degree to which commitments and theory "agree" (Beisbart, Betz, and Brun, 2021, 446).A theory accounts for a commitment, if the former entails the latter. Formally, a theory T *accounts* for a commitment $s \in C$ if and only if the commitment belongs to the dialectical closure of T, i.e., if and only if $s \in \overline{T}$. In order to arrive at a measure for account we can resort to compare commitments and theory as positions in our propositional framework and count sentences that differ with respect to their status (accepted, rejected, suspension). The following have a negative impact on the measure of account as they are a sign of misfit:

(Contradiction) commitments that are inconsistent with the theory, i.e., $c \in C$ and $\neg c \in \overline{T}$ (or vice versa)

(Contraction) commitments that are not entailed by the theory, i.e., $c \in C$ but $c \notin \overline{T}$

(Expansion) sentences entailed by the theory that are not part of the commitments, i.e., $t \in \overline{T}$ but $t \notin C$

The idea is that we go through every sentence (negated and unnegated) and penalise the cases of misfit (contradiction, contraction and expansion) between commitments and the dialectical closure of the theory. The fourth case is

(Agreement) commitments that are entailed by the theory, i.e., $c \in C$ and $c \in \overline{T}$,

and it does not receive a penalty. We introduce a penalty function d, and assume that the sentences s_i of the unnegated half of the sentence pool are indexed $s_1, s_2, \ldots, s_n$, where n is the size of the unnegated half of the sentence

pool. Let P and Q be positions on a dialectical structure.

$$d_{d_0,d_1,d_2,d_3}(P,Q,\{s_i,\neg s_i\}) = \begin{cases} d_3 & \text{if } \{s_i,\neg s_i\} \subset (P \cup Q) \text{ (contradiction)} \\ d_2 & \text{if } \{s_i,\neg s_i\} \cap (P) \neq \emptyset \\ & \text{and } \{s_i,\neg s_i\} \cap (Q) = \emptyset \text{ (contraction)} \\ d_1 & \text{if } \{s_i,\neg s_i\} \cap (P) = \emptyset \\ & \text{and } \{s_i,\neg s_i\} \cap (Q) \neq \emptyset \text{ (expansion)} \\ d_0 & \text{otherwise (agreement)} \end{cases}$$

d_3, d_2, d_1 and d_0 are penalties for contradiction, contraction, expansion, and agreement, respectively. By summing over all sentences, we define a weighted Hamming distance between positions P and Q.

$$D_{d_0,d_1,d_2,d_3}(P,Q) = \sum_{i=1}^{n} d_{d_0,d_1,d_2,d_3}(P,Q,\{s_i,\neg s_i\})$$

For account we specify the weights in the Hamming distance as follows: Agreement is not penalised, hence $d_0 = 0$. If a theory expands the commitments, is mildly penalised with $d_1 = 0.3$. If a theory fails to account for a commitment (contraction) the penalty is severe: $d_2 = 1$. Finally, contradictions are also severely penalised with $d_3 = 1$.

In a final step, the weighted Hamming distance of summed up penalties is normalised by the size of the unnegated half of the sentence pool (n), and handed over to a function G, that yields a monotonically decreasing function. G is defined as follows:

$$G(x) = 1 - x^2$$

and for commitments C and a theory T account is given by

$$A(C,T) = G(\frac{D_{0,0.3,1,1}(C,\overline{T})}{n})$$

$A(C,T)$ decreases with the number and penalties of misfitting sentences (expansion, contraction and contradiction). The maximal value that $A(C,T)$ can take is 1 corresponding to perfect agreement between commitments and the dialectical closure of the theory.

For $T = \{1\}$ with $\overline{T} = \{1, -2, 3, 4, 5, -6\}$, $C = \{3, 4, 5\}$, and $C' = \{2, 3, 4, 5\}$, the following values result:

$$A(C, T) = 0.983$$
$$A(C', T) = 0.948$$

T accounts for all elements in C, but it expands C with respect to $1, -2$, and -6, which is penalised. In addition, T contradicts C' with respect to -2, which results in a lower value for account.

Faithfulness operationalises the demand that the current commitments should respect the initial commitments. As both initial and current commitments are positions, we can again resort to a weighted Hamming distance to penalise deviations from the initial commitments. The penalties in the weighted Hamming distance for faithfulness are set as follows: $d_0 = 0$ (agreement), $d_1 = 0$ (current commitments expand the initial commitments), $d_2 = d_3 = 1$ (current commitments contract or contradict the initial commitments). Note that the penalties for account and differ with respect to d_1 for expansion. If the current commitments expand the initial commitments, there is no penalty, as the initial commitments are still respected. Analogous to account, the penalty function for faithfulness is normalised by the size of the unnegated half of the sentence pool n and turned into a monotonically decreasing function by G. Hence, we arrive at a measure for faithfulness:

$$F(C \mid C_0) = G(\frac{D_{0,0,1,1}(C_0, C)}{n})$$

Take for example $C_0 = \{3, 4, 5\}$ and $C = \{1, -2, 3, 4, 5, -6\}$. Even though C expands C_0 with respect to $1, -2$ and 6, $F(C \mid C_0) = 1$ is maximal because expansions are not penalised by faithfulness.

Systematicity models the demand in the informal account of RE that a theory should do justice to epistemic goals, especially theoretical virtues. Let T be a non-empty and dialectically consistent position, and hence the dialectical closure $\overline{T}$ is well-defined and non-empty as well. The measure for the systematicity of a theory is defined as follows:

$$S(T) = G(\frac{|T| - 1}{|\overline{T}|})$$

The expression inside of G relates the number of a theory's principles to the size of its content. If T is empty, $S(T)$ defaults to 0. The maximal value of S is 1.

Take again $T = \{1,7\}$ and $T' = \{3,7\}$ in the standard example. Both theories consist of two elements, but they differ substantially with respect to their contents: $\overline{T} = \{1,-2,3,4,5,-6,7\}$ and $\overline{T'} = \{3,7\}$. This results in $S(T) = 0.980$ and $S(T') = 0.750$.

Global Optima After having selected and specified three desiderata corresponding to three forces involved in RE, we need to aggregate them into an overall value for an epistemic state (C,T) given some initial commitments C_0 and handle trade-offs between desiderata. To this purpose Beisbart, Betz, and Brun (2021) introduce an *achievement function* which aggregates the desiderata.

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

Z is a convex combination of the desiderata measures since the weights for the desiderata are non-negative real numbers that sum up to 1:

$$\alpha_A + \alpha_S + \alpha_F = 1$$

The weights determine how trade-offs between desiderata are dealt with. If, for example, systematicity (α_S) has more weight than account (α_A), then more systematic theories with limited account may still be preferable over complex theories with better account. For other weightings, trade-offs may turn out differently.

Assume that a dialectical structure τ and weights α_A, α_S and α_F are given. A *global optimum* relative to some initial commitments C_0 is an epistemic state (C,T) such that the achievement function $Z(C,T \mid C_0)$ is maximal. This means that there is no other epistemic state that performs strictly better with respect to Z than (C,T). From the assumption that the sentence pool is finite, we can conclude that there is at least one global optimum. However, due to equally well-performing epistemic states, there may be multiple global optima relative the same set of initial commitments.

RE States Have we found an RE state if an epistemic state is a global optimum according to the achievement function relative to some initial commitments? Beisbart, Betz, and Brun (2021, 449) answer in the negative and propose additional optimality conditions on epistemic states taken from the literature on RE.

(CCT) The commitments and the theory are consistent with each other.

(FEA) The theory fully and exclusively accounts for the commitments.

Dialectical compatibility operationalises (CCT) in the model. Commitments and theory are consistent with each other if there is a complete and consistent position that extends both. Formally, (FEA) means that the commitments and the closure of the theory of an epistemic state (C, T) coincide, i.e., $C = \overline{T}$. Note that (FEA) is a more ambitious condition as (CCT) as the former implies the latter. Consequently, we can distinguish between two kinds of RE states depending on whether a global optimum satisfies (CCT) or (FEA).

(RE State) An epistemic state (C, T) is an *RE state* (relative to initial commitments C_0) if and only if (i) it is a global optimum according to the achievement function Z, and (ii) the commitments and the theory are consistent with each other (CCT).

(Full RE State) An epistemic state (C, T) is a *full RE state* (relative to initial commitments C_0) if and only if (i) it is a global optimum according to the achievement function Z, and (ii) the theory fully and exclusively accounts for the commitments (FEA).

The relativisation of (full) RE states to the inputs, in particular to the set of initial commitments C_0, rests on the assumption that other operationalised components of RE are given and held fixed, e.g., the dialectical structure (Beisbart, Betz, and Brun, 2021, 464). In view of the upcoming ensembles of simulations that vary more than just initial commitments, and to stress the relative character of RE results, it may be better to report results of RE relative to a dialectical structure $\tau = (\mathcal{A})$, initial commitments C_0, and the weighting $(\alpha_A, \alpha_S, \alpha_F)$. Still, this runs on the assumption that the penalties for measures in the achievement function, and the achievement function itself are given and held fixed.

RE Process Informal account also involves a process of mutual adjustments between commitments and theory, which is formally modelled as follows: The initial state is given by the initial commitments C_0 and theory T_0, which is assumed to be empty.

(Theory Adjustment) Given the current commitments C_i, the agent searches for a new theory T_{i+1} that maximises the achievement function

$$Z(C_i, T_{i+1} \mid C_0).$$

There may be several maxima due to ties. If T_i is among them, $T_{i+1} = T_i$. Otherwise T_{i+1} is chosen randomly among the maxima. Adopting the new theory results in an updated epistemic state (C_i, T_{i+1}).

(Commitment Adjustment) Given the current theory T_{i+1}, the agent searches for a new set of commitments C_{i+1} that maximises the achievement function

$$Z(C_{i+1}, T_{i+1} \mid C_0).$$

There may be several maxima due to ties. If C_i is among them, $C_{i+1} = C_i$. Otherwise C_{i+1} is chosen randomly among the maxima. Adopting the new commitments results in an updated epistemic state (C_{i+1}, T_{i+1}).

The agent consecutively applies the rules for theory and commitment adjustments until there are no further changes. If the following condition is met, the process terminates with a epistemic state (C_{i+1}, T_{i+1}), that is called *fixed point*.

(Stopping Rule) The process of equilibration terminates if $(C_{i+1}, T_{i+1}) = (C_i, T_i)$.

Beisbart, Betz, and Brun (2021) study four cases in the standard example, which are given by four different initial commitments. The weights are set to $(\alpha_A, \alpha_S, \alpha_F) = (0.35, 0.55, 0.10)$. The results are collected in Table 7.1.

case	initial commitments C_0	fixed points (C, T)	global optima (C, T)
A	$\{3, 4, 5\}$	$(\{1, 3, 4, 5, -6, -2\}, \{1\})$	$(\{1, 3, 4, 5, -6, -2\}, \{1\})$
B	$\{2, 3, 4, 5\}$	$(\{1, 3, 4, 5, -6, -2\}, \{1\})$	$(\{1, 3, 4, 5, -6, -2\}, \{1\})$
C	$\{3, 4, 5, 6, 7\}$	$(\{1, 3, 4, 5, -6, -2\}, \{1\})$ $(\{2, 5, 6, 7, -4, -1\}, \{2\})$	$(\{1, 3, 4, 5, -6, -2\}, \{1\})$ $(\{2, 5, 6, 7, -4, -1\}, \{2\})$
D	$\{3, 4, 5, -6, 7\}$	$(\{1, 3, 4, 5, -6, -2\}, \{1\})$	$(\{1, 3, 4, 5, -6, -2\}, \{1\})$

TABLE 7.1: RE outputs in four cases (initial commitments) in the standard example. Multiple outputs arise due to random choices in equilibration processes (fixed points), or due to equally best performing states (global optima).

All outputs of Table 7.1 happen to satisfy (FEA), and hence qualify as full RE states. It is important to note that things may not turn out so nicely in every example. Fixed points and global optima that are reachable from a specific set of initial commitments can come apart: Some fixed points are not

globally optimal, and some global optima are not reachable by equilibration processes that set out from that specific set of initial commitments.

Differences arise from the contrast of global optimisation and the process of equilibration as alternating, *semi-global* optimisations. In every step, all candidates are considered, i.e., all minimally consistent positions for commitments, and all dialectically consistent positions for theories, respectively. The other part of the epistemic state is held fixed. Consequently, faithfulness does not make a difference in the achievement function for theory adjustment, as it does not depend on the theory. In turn, systematicity does not have an influence on the achievement in commitment adjustment steps.

There is a series of interesting remarks that ensues from the distinction between fixed points reached by a process of alternating semi-global optimisation, and global optima according to the achievement function. First of all, the model allows to study the dynamic and the static aspects of RE separately, as it produces outputs from the process of equilibration as well from the global evaluation of epistemic states.

As a consequence of the semi-global optimisation, the fixed point resulting from an RE process may not be a global optimum according to the achievement function, and hence, not qualify as an RE state. Moreover, even global optimisation does not guarantee that the global optima satisfy the additional requirements (CCT) or (FEA). This renders both aspects of the model "imperfect" in the sense of Elgin (1996, 4): There is a (process-)independent "criterion of correctness", namely the conditions for (full) RE states, but neither equilibration process nor global optimisation guarantee that the outputs meet these requirements.

Finally, both fixed points and global optima satisfy the fourth condition on RE states from Section 2.2.1: They are optimal among available epistemic states ("as reasonable as any available alterntive"), but the domain of what is available differs. And the difference is much more pronounced than what technically correct term "semi-global" would lead us to expect. To illustrate the difference, think of epistemic states (C, T) as cells on a appropriately sized, possibly non-square, chess-board. The unbounded, globally optimising agent can overview the entire board at once, while the other agent can evaluate but a single row or column per adjustment step. In the standard example, there are 1,163 dialectically consistent positions for theory candidates, and 2,186 minimally consistent positions for commitments that are taken into consideration in an alternating fashion during the process of equilibration. In contrast, global optimisation effectively evaluates $1{,}163 \times 2{,}186 = 2{,}542{,}318$

epistemic states. Equilibration processes, which tend to terminate after a few steps, cannot catch up with this difference.

7.1.3 Comparison to Informal RE

Lets take a step back and compare the formal model of RE to the elaborate, informal account of RE that I presented in Chapter 2. Table 7.2 lists the components of elaborate informal accounts of RE and the formal model.

informal account	**formal model**
commitments	dialectical position C
input commitments	initial commitments C_0
theory	dialectical position T
position of an agent	epistemic state (C, T)
background	dialectical structure τ
fit	account $A(C, T)$
doing justice to epistemic goals	systematicity $S(T)$
respecting the input	faithfulness $F(C \mid C_0)$
securing independent credibility	
aggregation	achievement function $Z(C, T \mid C_0)$
trade-offs	weights $\alpha_A, \alpha_S, \alpha_F$
process of equilibration	adjustment rules
endpoint of process	fixed point
equilibrium state	global optimum + (CCT) or (FEA)

TABLE 7.2: Informal RE components and demands with their formal counterparts.

The most notable discrepancy is the lack of a formal counterpart for the weakly foundationalist demand of securing independent credibility. Independent credibility could be implemented straightforwardly in the model by assigning numerical values to commitments. However, this would increase the complexity of the model with a substantial amount of free parameters, and open a series of very interesting question that go beyond the scope of this project. Is a independent credibility of a commitment a subjective probability? How do we determine a numerical value that represents independent credibility? Is independent credibility subject to change during an RE

process? What would be the adjustment rules? Excluding independent credibility for the moment is not a severe shortcoming as modelling the desideratum of respecting initial commitments by faithfulness already puts a non-coherentist demand on a position that "ties" the commitments to the input. Moreover, given our focus on theoretical virtues that relate to coherence, the exclusion is defensible.

As a consequence of the fairly close correspondence, we can reuse the schematics of the informal account (Figure 2.1 in Chapter 2) to portray the formal model in the same way in Figure 7.2).

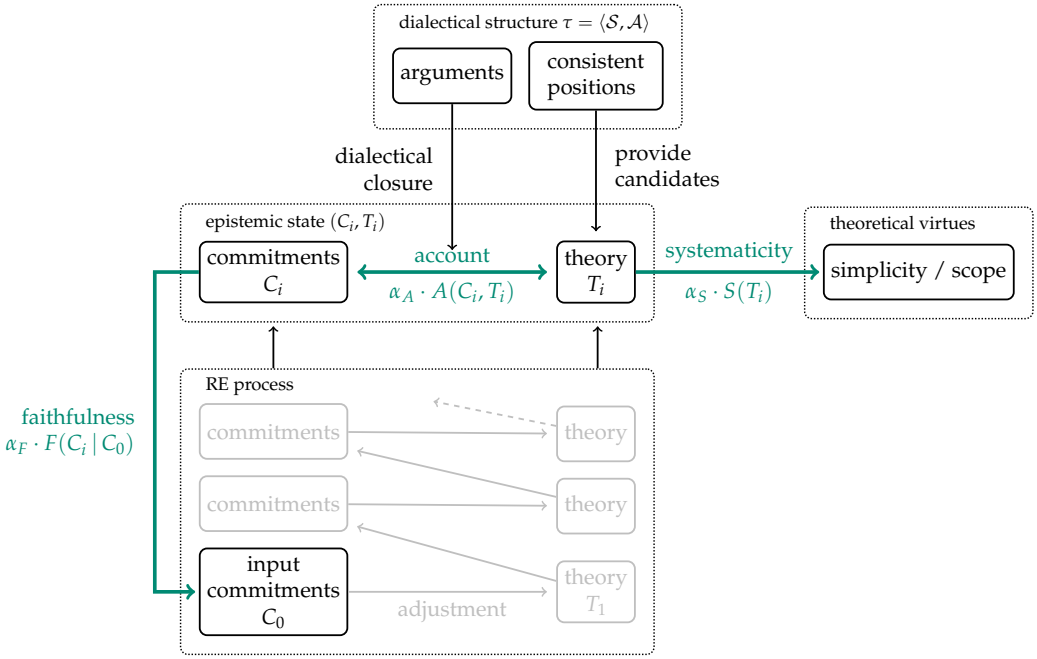


FIGURE 7.2: The formal model of RE illustrated in the same schematics as the informal account.

In conclusion, we see that the formal model represents many key components of RE: epistemic states distinguish between commitments and a theory, RE states and processes are guided by considerations of fit, respecting the input and doing justice to theoretical virtues. Note that formal model goes far beyond completely "narrow" RE, which aims at establishing fit between commitments and theory. We can interpret dialectical structures to take up the role of the background of inquiry (Beisbart, Betz, and Brun, 2021, 460). Dialectical structure provide theory candidates that are required to be dialectically consistent. The dialectical closure of a theory occurs in account, and hence the background also facilitates fit between commitments and theory.

7.2 Discussing the Model

This section addresses the question where theoretical virtues come into play in the formal model. It turns out, that it incorporates a wide range of virtues. Moreover, it covers the coherential virtues that I presented in the framework of virtue-based coherence in Chapter 5.

7.2.1 Theoretical Virtues Implemented in the Model

The formal model of RE incorporates theoretical virtues as desiderata and requirements for RE states. Some of these virtues concern the theory on its own, the relation of a theory to a set of commitments, or the theory in view of the background. Here, I will also include what I have called *hybrid* virtues earlier, i.e., virtues that involve theories, but may not be purely theoretical. In our case, hybrid virtues will be virtues of a theory in relation to some commitments that are held in an epistemic state (C, T), or in relation to the dialectical structure τ in the background. Features of sets of commitments, e.g., minimal consistency or faithfulness to the initial commitments are not covered, as they are not theoretical virtues. Consequently, the model exemplifies my adaptation of Douglas' systematisation of theoretical virtues (2013) from Section 4.1.

Table 7.3 summarises the theoretical virtues that are present in the formal model during theory adjustment steps in equilibration processes, as well as the additional requirements on global optima to qualify as full RE states.[6] We

	pure	**hybrid**
requirement	minimal consistency	dialectical consistency (CCT) (FEA)
desideratum	simplicity	account scope

TABLE 7.3: Classification of theoretical virtues implemented in the formal model according to the distinctions between pure and hybrid virtues, as well as requirements and desiderata. (CCT) and (FEA) are additional requirements for global optima to qualify as (full) RE state.

[6]Commitment adjustment steps are also guided by virtues that apply to commitments themselves (minimal consistency), in relation to the initial commitments (faithfulness), and in relation to theories and background (account).

could further distinguish hybrid virtues, which pertain to the theory in relation to the background (dialectical consistency, scope), from those, which relate a theory to the commitments given the background (account, CCT, FEA). I present them in more detail in the following paragraphs.

Consistency The theoretical virtue of consistency is implemented into the model in various places. We can think of it as a three-layered structure: minimal, dialectical and mutual consistency build upon each other:.

Minimal consistency is both a requirement for theories as well as commitments on their own, meaning that no position, which contains both a sentence and its negation, is admissible. We can interpret minimal consistency for theories to reflect the theoretical virtue of *internal consistency* in a crude manner. Minimal consistency does not take into account inferential relations provided externally by the dialectical structure, but it requires that a position does not contain flat contradictions, which is a completely internal affair of a position. Admittedly, this is a very weak requirement. However, we can see it as absolutely basic condition of rationality, at least if we accept the classical law of non-contradiction.

Next, theories (or candidate positions for theories) are required to be dialectically consistent. Dialectical consistency is a stronger requirement than minimal consistency as the latter is implied by the former. As dialectical consistency is external affair of a position in relation to the dialectical structure, which can be understood as the background of an RE setting (Beisbart, Betz, and Brun, 2021, 460), dialectically consistent theories exhibit the virtue of *external consistency*.

The third layer of consistency in RE that is relevant to the definition of (full) RE states, i.e., epistemic states that are global optima according to the achievement function satisfying (CCT) or (FEA). (CCT), which is implied by (FEA), requires that commitments and the theory of an epistemic state are consistent with each other, I will call it *mutual* consistency. Formally, mutual consistency is implemented as *dialectical compatibility*. Two positions are dialectically compatible if and only if they have a complete and consistent extension in common. Consequently, mutual consistency is related to dialectical consistency, but it now involves commitments as well as the theory. Mutual consistency is again stronger than dialectical consistency as the former implies the latter (for both theory and commitments). Note that the converse does not hold. Even if commitments and the theory of an epistemic state are

dialectically consistent on their own, they may still fail to be consistent with each other.

Mutual consistency spells out the idea that commitments and theory cohere in a very weak sense in that they are consistent with each other. Hence, mutual consistency can be seen as the theoretical virtue of consistency that is commonly considered as a condition of coherence. The strive to incorporate this virtue into a coherent outcome of RE motivates the additional optimality condition (CCT) that is required of a global optimum to qualify as an RE state in the formal model.

Can consistency be traded off against other theoretical virtues in the formal model of RE? The answer depends on the layer of consistency. Dialectical consistency (and by implication minimal consistency, too), is a necessary requirement of theories that cannot be traded-off against other virtues in the formal model. In contrast, mutual consistency, which is recommended by account and a dialectically consistent theory, can be traded off for specific weightings that give strong preference to faithfulness.[7]

Minimal consistency is a purely theoretical virtue, dialectical and mutual consistency are hybrid as they relate theory and background, or commitments, theory and background, respectively.

Account Account measures to which extent the former are inferrable from the latter. Thus, account comes in degrees, and it is a hybrid virtue, as it pertains to the theory in relation to the commitments given the background of the dialectical structure, that provides the inferential relations between them. Arguably, account is reminiscent of the theoretical virtue, which Kuhn (1977, 357) calls *accuracy*, but stripped of its empiricalness (see Section 5.2).

Account is weighted in the achievement function, and hence it can be traded-off against faithfulness and systematicity. The additional requirement (FEA) for global optima to qualify as full RE states demands that the theory T fully and exclusively accounts for the commitments C ($\overline{T} = C$). The reason to require (FEA) of full RE states is that any deviation from full and exclusive account is an epistemic shortcoming, especially from the viewpoint of coherence. If (FEA) does not hold, there are inconsistencies, unaccounted for commitments, or consequences of the theory that are missing among the commitments. Such cases are fatal or at least detrimental to coherence in terms of establishing inferential relations.

[7] We can observe this in results from simulations presented in Section 10.4. If fixed point or global optimum commitments are inconsistent, they immediately fail to satisfy (CCT).

Systematicity Table 7.3 does not list systematicity as a separate virtue because I suppose that it can be split into the virtues of (syntactic) simplicity and scope. Recall that a theory's systematicity is measured by

$$S(T) = G(\frac{|T| - 1}{|\overline{T}|}).$$

The numerator of the expression inside of G, $|T| - 1$, is a crude measure of a theory's complexity that increases in terms of numbers of elements in a theory. The application of function G turns the measure of complexity into a measure of simplicity. Having fewer elements in a theory is beneficial to the measure of systematicity. In turn, the denominator $|\overline{T}|$ brings the size of the dialectical closure of the theory into play. This captures the scope of a theory and the measure of systematicity treats broad scope as an attractive feature of a theory.

Simplicity is a purely theoretical virtue, and as scope depends on the theory and the dialectical structure in the background, I count it among the hybrid theoretical virtues.

The measure for systematicity S is weighted in the achievement function Z. Consequently, systematicity can be traded-off against the other components in Z, namely account and faithfulness. By considering the ratio of complexity and scope, i.e., the fraction of inside of G in the measure of systematicity, we can observe an additional trade-off between complexity (simplicity) and scope. If a theory T' is more complex (less simple) than a theory T it may still be the case that $S(T') > S(T)$ if T' has broader scope than T. I relegate the formal analysis of this trade-off to the next chapter. For now, it suffices to note that systematicity recommends theories that strike a balance between being simple theories and having a broad scope.

7.2.2 Comparison to Virtue-Based Coherence

How does the formal model of RE by Beisbart, Betz, and Brun (2021) in the framework of TDS relate to the BRT-inspired framework for virtue-based coherence in Chapter 5? Let $B = L \cup I$ be a belief base consisting of literals (L) and inferential beliefs (I). It turns out that we can translate the epistemic state represented by B into the TDS framework. A dialectical structure τ consists of a sentence pool $\mathcal{S}$ and a set of deductively valid arguments $\mathcal{A}$. The sentence pool corresponds to the set of literals $\mathfrak{L}$, which is assumed to be finite and closed under negation. The inferential relations of I can be converted

to deductive arguments, by designating one disjunct as conclusion and the others as premises. As $\mathcal{A}$ consists of all arguments, it is not important, which disjuncts are treated as a premises. Finally, L corresponds to a position on a dialectical structure.

There is a discrepancy that stands in the way of translation in the other direction. The sentence pool of TDS represents natural-language sentences, while $\mathfrak{L}$ consists of literals of a formal language, i.e., atomic propositions and their negations. Thus, the translation works in the other direction if we assume that the sentences in the sentence pool are literals (which might be what we do in practice). In fact, Beisbart, Betz, and Brun (2021, 460) propose to choose the sentence pool to include atomic sentences (and to exclude complicated principles) in concrete applications in view of the worry that their measure of systematicity is too idealised. In this case, a position P on a dialectical structure τ gives rise to a belief base $B = L \cup I$, where L consists of the literals adopted in P, and I results from all inferential relations imposed by the arguments of τ.

Thus, for appropriately chosen sentence pools, we can translate between the frameworks of TDS and BRT. This opens up an interesting line of future research as the relation between TDS and BRT has yet not been explored, as far as I am aware. Investigation in this direction might prove to be insightful for RE, as BRT also offers a rich set of tools and results, in particular belief changing operations and representation theorems.

In addition, we can observe that the list of virtues discussed in the formal model and in the framework for virtue-based coherence coincide. The reader is advised to take the close correspondence between the virtue-based coherence for RE and the formal model of RE with a grain of salt, as I may be biased towards the formal model. Genealogically, I was first acquainted with the formal model of Beisbart, Betz, and Brun (2021), and then developed my account of virtue-based coherence. This results in a high degree of shared terminology and close correspondence between the selection and specification of virtues in these frameworks.

Still, I think that the formal approach of virtue-based coherence enjoys sufficient motivation to be a worthwhile addition to the formal model, as it shows that the ideas voiced by Beisbart, Betz, and Brun (2021) can be formalised consistently in a different framework, and it lends independent support to include theoretical virtues. The inferential relations in TDS are given by deductive arguments. I argued in Chapter 5.1 that spelling out coherence as consistency and support from residuals by solely deductive inference is

unsatisfactory to some extent. The addition of virtues to a deductive framework serves to arrive at a more substantive notion of coherence without having to include non-deductive inferential relations.

Now, we have seen, that the formal model of Beisbart, Betz, and Brun (2021) does an excellent job of incorporating coherential virtues, and it arrives at a fully operational configuration of theoretical virtues for RE. The achievement function Z takes care of the aggregation of the weighted measures on ratio scales. In contrast to weaker (i.e., less informative) ordering relations, they effectively escape Okasha's (2011) application of Arrow's theorem to theory choice by means of virtues (Chapter 6) .

Chapter 8

Analysing the Formal Model

The formal model of RE provided by Beisbart, Betz, and Brun (2021) implements a fully operational configuration of theoretical virtues. In particular, the formal model offers the means to weigh desiderata. Are some weightings more plausible than others? Or are agents free to assign weights as they like?

Formalisation allows for exploration by analytical means. Beisbart, Betz, and Brun (2021, 450) state and prove basic propositions about the formal model, especially about equilibration processes and the relation between fixed points and full RE states. In this chapter, I aim to add further results that contribute to the formal model's fruitfulness even before we begin to produce and explore data. The focus rests on the role of theoretical virtues in the formal model of RE, and in particular, on the weightings and their aggregation in the achievement function. Furthermore, some findings concern the robustness of the formal model with respect to weightings. This is an important step towards devising weights for a plausible configuration of theoretical virtues that can be used to address objections to RE.

Some of the following results are rather technical and reached through cumbersome proofs. Nonetheless, they deepen our understanding of the inner workings of the model. A formal model should not be a black box if it is supposed to shed light on RE. Moreover, most of them prove to be useful for producing data or interpreting results later.

The chapter is structured as follows: In a preliminary remark, I introduce ternary plots that allow for an elegant visualisation of the space of configurations. In Section 8.1, I present various fruitful analytical findings. The model's robustness with respect to configurations is analysed in Section 8.2. In Section 8.3, a series of plots illustrates some of the results.

Ternary Plots We can think of the weights $(\alpha_A, \alpha_S, \alpha_F)$ as a point in three-dimensional space, in particular $\mathbb{R}^3$. I will call $(\alpha_A, \alpha_S, \alpha_F)$ a *configuration of weights*, and I assume that the possible confusion with a configuration of

theoretical virtues is manageable. A configuration of weights is part of the weighting in a configuration of theoretical virtues. The set of configuration of weights is a plane in three-dimensional space due to the boundary condition $\alpha_A + \alpha_S + \alpha_F = 1$, as illustrated in Figure 8.1.

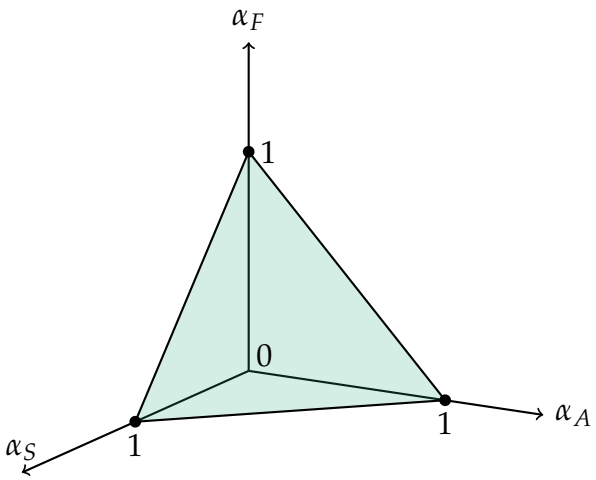


FIGURE 8.1: The space of weight configurations (green) is a plane in three-dimensional space satisfying $\alpha_A + \alpha_S + \alpha_F = 1$.

Unfortunately, three-dimensional plots are arduous in production, deceitful in interpretation, and the fact that the space of a configuration is a plane anyway, speaks in favour of opting for a two-dimensional approach. An easy solution is to project the space of weight configurations down onto the $\alpha_A\alpha_S$-plane, by dropping the α_F value. The result is depicted in Figure 8.2. Note that the value for α_F can be reconstructed by $\alpha_F = 1 - (\alpha_A + \alpha_S)$.

The fact that the weights sum up to 1 allows for a more elegant two-dimensional depiction without having to drop the the third coordinate: ternary plots, such as Figure 8.3. Here is a quick guide on how to read data from a ternary plot: Every weight is maximal ($= 1$) in a corner of the triangle (α_A: top, α_S: bottom left, α_F: bottom right) and minimal ($= 0$) on the opposite edge, respectively. Between a corner and its opposite edge, the weights decrease in a linear fashion. In the case of α_A, the horizontal lines in Figure 8.3 represent configurations with equal values for α_A.

8.1 Fruitfulness

Perfect States An epistemic state is "perfect" if and only if it maximises the achievement function Z with a value of 1. In this case the following proposition holds:

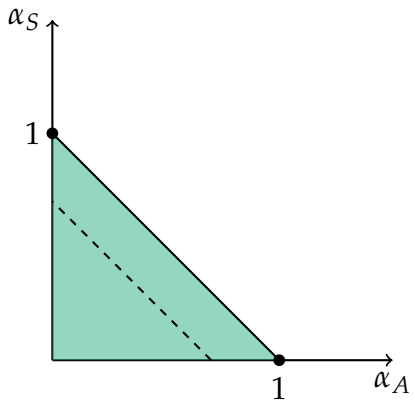


FIGURE 8.2: Projection of the space of weight configurations onto the plane of α_A and α_S. Diagonals (e.g., the dashed line) represent configurations that have a constant value for α_F that increases towards the origin.

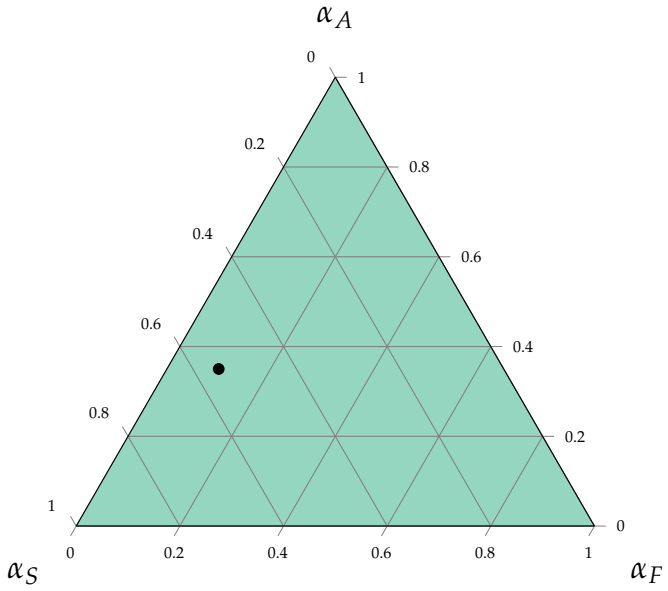


FIGURE 8.3: The space of weight configurations in a ternary plot. The black dot marks the default configuration of weights in the formal model: $(\alpha_A, \alpha_S, \alpha_F) = (0.35, 0.55, 0.1)$.

Proposition 2. *Assume that a dialectical structure τ and some initial commitments C_0 are given. Let (C, T) be a global optimum (relative to C_0) according to the achievement function Z specified with a configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$ such that $Z(C, T \mid C_0) = 1$. Then (C, T) is a global optimum according to Z' specified for every configuration $(\alpha'_A, \alpha'_S, \alpha'_F)$, and $Z'(C, T \mid C_0) = 1$.*

Proof. $Z(C, T \mid C_0) = 1$ holds if and only if every measure in Z maximised. Thus, $A(C, T) = S(T) = F(C \mid C_0) = 1$. This implies for every configuration of weights $(\alpha'_A, \alpha'_S, \alpha'_F)$:

$$Z'(C, T \mid C_0) = \alpha'_A \cdot A(C, T) + \alpha'_S \cdot S(T) + \alpha'_F \cdot F(C \mid C_0) = \alpha'_A + \alpha'_S + \alpha'_F = 1$$

As 1 is the maximal value that Z' can take, (C, T) is a global optimum according to Z'. □

Perfect states involve no trade-offs between the measures for epistemic desiderata. Hence, the configuration of weights, which guides the trade-offs, is irrelevant.

Systematicity I discussed the theoretical virtues of syntactical simplicity and scope as important ingredients to spell out coherence in a deductive framework (Chapter 4). The formal model of Beisbart, Betz, and Brun (2021) implements these virtues in the measure for systematicity $S(T)$ (see Chapter 7.2). Here, I remark on two points.

First, the current measure of systematicity S has a shortcoming. If a theory contains exactly one element, say $T = \{1\}$, the measure is maximised ($S(T) = 1.0$) because the numerator ($|T| - 1$) is 0. In this case the denominator ($|\overline{T}|$), i.e., the number of sentences that can be inferred from the theory given the dialectical structure, is not relevant. This amounts to the failure of differentiating singleton theories according to their scope (the size of their dialectical closure). This leads to a bias of the formal model towards epistemic states with a singleton theory, irrespective of the theory's scope.

Second, the numerator $|T| - 1$ denotes a score for the theory's syntactical complexity, and the denominator $|\overline{T}|$ measures the scope. As a fraction, they model a trade-off between syntactic complexity and scope.

Formally, we can determine this trade-off more precisely: Let T, T' be non-empty, dialectically consistent positions from a dialectical structure with n unnegated sentences in its pool, such that $m = |T|$ and $k = |\overline{T}|$ (and analogously for T'). Assume $m, m' > 1$ to exclude the pathological case of singleton theories. Note that k corresponds to the score for scope and $m - 1$ to the

score for syntactic complexity. As $T \subseteq \overline{T}$, we have $m \leq k$. Assume that T' has more elements than T, i.e., $m < m'$, resulting in

$$1 < m, k < n \text{ and } k \geq m,$$

$$1 < m', k' \leq n \text{ and } k' \geq m'.$$

In which circumstances would we be willing to say that T' is more systematic than T although the former is more complex? A plausible condition is the following: T' is more systematic than T if and only if the relative increase in scope is greater than the relative increase in syntactic complexity by adopting T' rather than T. We can write down the condition formally:

$$\frac{m' - 1}{m - 1} < \frac{k'}{k}$$

This is equivalent to

$$\frac{m' - 1}{k'} < \frac{m - 1}{k}.$$

The application of the monotonically decreasing function G to both sides reverses the inequality:

$$G\left(\frac{m' - 1}{k'}\right) > G\left(\frac{m - 1}{k}\right),$$

and by substitution, we have

$$G\left(\frac{|T'| - 1}{|\overline{T'}|}\right) > G\left(\frac{|T| - 1}{|\overline{T}|}\right).$$

Finally, we arrive at

$$S(T') > S(T),$$

which shows that T' is indeed more systematic than T according to S. As the implications used in above line of thought also work in the other direction, equivalence ensues. Increasing the syntactic complexity of a theory is beneficial to its systematicity if and only if the increase in syntactic complexity is surpassed by the increase in scope.

Perhaps, one might want to have more control of the trade-off between complexity and scope with additional weights. One way to achieve this straightforwardly, is to split the measure of systematicity into two separate measures for simplicity and scope, and include them in the achievement function with corresponding weights. The normalised measures, which are

inspired by the measures developed in the framework for virtue-based coherence in Section 6.3, could look something like this:

$$Simplicity(T) = G\left(\frac{|T| - 1}{n - 1}\right)$$
$$Scope(T) = G\left(\frac{n - |\overline{T}|}{n - 1}\right)$$

Note that separating simplicity and scope would take care of the shortcoming in the measure of systematicity with respect to singleton theories.

Faithfulness A proposition of Beisbart, Betz, and Brun (2021, 450) states that "for extreme parameter choices (i.e., $\alpha_A = 0$ or $\alpha_A = 1$), every equilibration fixed point is a global optimum." Similarly, we can prove the following for an extreme choice of faithfulness:

Proposition 3. *Assume that a dialectical structure τ and some initial commitments C_0 are given. If $\alpha_F = 0$, then every global optimum (relative to C_0) consists of a singleton theory and its dialectical closure as commitments (irrespective of the other weights α_A and α_S). Furthermore, every global optimum is a full RE state.*

Proof. Assume that $\alpha_F = 0$, $\alpha_A, \alpha_S \neq 0$, and $\alpha_A + \alpha_S = 1$. Thus,

$$Z(C, T \mid C_0) = \alpha_A \cdot A(C, T) + \alpha_S \cdot S(T) + 0 \cdot F(C \mid C_0)$$

This means that the Hamming distance to the initial commitments C_0 (inside of faithfulness $F(C \mid C_0)$) does not have a bearing on the optimality of an epistemic state (C, T). We have noted earlier, that $S(T) = 1$ if and only if T is a singleton theory, and indeed, there is at least such a dialectically consistent position.[1]

If we take the dialectical closure of a singleton theory T as commitments ($C = \overline{T}$), account is maximal: $A(C, T) = 1$. Consequently, we have

$$Z(C, T \mid C_0) = \alpha_A \cdot A(C, T) + \alpha_S \cdot S(T) = \alpha_A \cdot 1 + \alpha_S \cdot 1 = 1 \tag{8.1}$$

It follows that the epistemic state (C, T) is a global optimum since 1 is the maximal value of Z. This holds for every epistemic state consisting of a singleton theory and its closure as commitments.

[1] This follows from the basic assumption that a dialectical structure has at least one complete and consistent position. Thus, there are at least n singleton theories, where n is the number of elements in the unnegated half of the dialectical structure's sentence pool. The maximum number of singleton theories is $2 \cdot n$, but it may be lower if some of them are dialectically inconsistent.

Since both account A and systematicity S are maximised by such states, any other epistemic state, which consists of a theory with more than one element or commitments that are not perfectly in agreement with the theory, performs worse according to the achievement function.

Since $A(C, T) = 1$ holds for those global optima, the theory full and exclusively accounts for the commitments (FEA). Hence, they qualify as full RE states. Furthermore, Equation (8.1) shows that the specific weights of α_A and α_S do not matter for obtaining this result. □

With $\alpha_F = 0$, the epistemic agent can "leap" (choose commitments and a theory simultaneously) without having to take the initial commitments into consideration for global optimisation. In contrast, the first step in an equilibration process evaluates theory candidates in view of the initial commitments. Even though faithfulness is nullified in the achievement function, candidate theories are evaluated with respect to how well they account for the initial commitments. Hence, the first step in an equilibration still involves a tie to the initial commitments irrespective of the weight for faithfulness.

What lesson can we draw from this, especially with reference to theoretical virtues in the formal model of RE? The "conservative" pull of faithfulness sometimes works "against" the theoretical virtues that are implemented with account and systematicity (see Chapter 10 and Chapter 11). So, one may be tempted to give up faithfulness completely in order to render the outcomes of RE more virtuous. However, the present results suggest that this is not a good idea. The many globally optimal states are extremely simplistic, yet they still satisfy the conditions for full RE states. Moreover, it also renders trade-offs between account and systematicity irrelevant to global optima.

A Linear Model Variant The monotonically decreasing function

$$G(x) = 1 - x^2$$

turns the penalties into measures of account, systematicity and faithfulness. So far, I did not motivate the use of a quadratic function instead of a simpler linear function, and neither do Beisbart, Betz, and Brun (2021):

$$G(x) = 1 - x$$

This is relevant to the discussion of theoretical virtues in RE, as the choice of function G is part of specifying the aggregation of theoretical virtues in the model.

A simple but plausible motivation for choosing a quadratic function is the following: Small deviations from the optimum should not be penalised as much as big ones. For example, the loss in systematicity when going from a theory with two principles to one with three principles should be less severe than going from three to four principles. This is illustrated in Figure 8.4. With growing x (penalties) the linear function $1 - x$ decreases by a constant amount. In contrast, the quadratic function $1 - x^2$ penalises less severely near the optimal value on the far left side for small penalties x and more drastically (steeper slope) for high penalties on the right side.

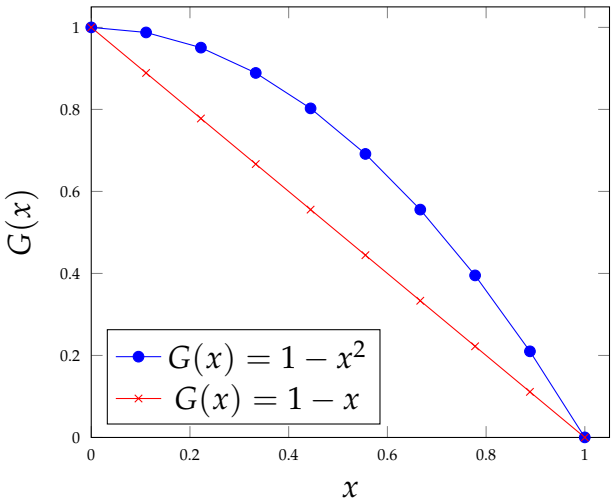


FIGURE 8.4: Comparison of values from a quadratic and a linear function.

A *linear model variant*, is a model that differs from the model provided by Beisbart, Betz, and Brun (2021) only in that the measures involve the linear function G instead of the quadratic one. This variant exhibits a *tipping line* in ternary plots that marks off configurations of weights that lead to drastically different behaviour.

We can characterise the tipping line as an equation that relates α_A and α_S:

$$\alpha_A = \frac{1 - \alpha_S}{2} \tag{8.2}$$

The boundary condition $\alpha_A + \alpha_S + \alpha_F = 1$ allows us to rewrite Equation 8.2 in an even simpler form:

$$\alpha_A = \alpha_F$$

Consequently, the tipping line splits the space of weight configuration in two regions $\alpha_A < \alpha_F$ and $\alpha_A > \alpha_F$. For the latter region, where account receives more weight than faithfulness, we have some interesting analytical results.

Proposition 4. *Assume that a dialectical structure τ and some initial commitments C_0 are given. Moreover, assume $\alpha_A > \alpha_F$ for a configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$ in a linear model variant. Then all global optima (relative to C_0) according to the achievement function specified by the configuration of weights are full RE states.*

Proof. Intuitively, $\alpha_A > \alpha_F$ means that account trumps faithfulness, which allows to select commitments ignoring faithfulness so that they are fully and exclusively accounted for by a theory.

Assume that an epistemic state (C, T) is a global optimum according to the achievement function Z given some initial commitments C_0 and a configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$ such that $\alpha_A > \alpha_F$. We need to show that (C, T) is a full RE state, i.e., that T fully and exclusively accounts for C (FEA), or equivalently, $A(C, T) = 1$.

For a proof by contradiction, assume that

$$A(C, T) = G(\frac{D_{0,0.3,1,1}(C, \overline{T})}{n}) < 1,$$

which holds only if $D_{0,0.3,1,1}(C, \overline{T}) > 0$. In other words, there is at least one sentence s (negated or unnegated) for which there is a positive contribution to the Hamming distance. In particular, we have the following cases:

i) $\overline{T}$ extends C with respect to s: $+0.3$

ii) $\overline{T}$ contracts C with respect to s: $+1$

iii) $\overline{T}$ and C contradict each other with respect to s: $+1$

Consider the impacts to the contributions to the Hamming distances for account and faithfulness of changing C with respect to s, yielding new commitments C' in Table 8.1. Note that systematicity is not affected by changing the commitments.

Let me explain how to read, for example, the third row of this table. The worst case for faithfulness (i.e., the highest valued penalty) is the following: There is $s \in \overline{T}$ such that $s \notin C$ and $\neg s \in C_0$. Consequently, adding s to C (yielding C') introduces a contradiction to the initial commitments C_0, which is penalised $(+1)$ by $D_{0,0,1,1}(C_0, C')$ in the measure of faithfulness. But by assumption, $s \notin C$ and $\neg s \in C_0$, implying that $D_{0,0,1,1}(C_0, C)$ also contributes

a penalty (+1), since C_0 extends C with respect to s. Hence, the difference of contributions to the Hamming distances is 0. In turn, adding the missing element to C takes care of the expansion of C by $\overline{T}$ with respect to s, which contributes a penalty (+0.3).

change	**account** $d_{0,0.3,1,1}(C', \overline{T}, \{s, \neg s\})$ $-d_{0,0.3,1,1}(C, \overline{T}, \{s, \neg s\})$	**faithfulness** (worst case) $d_{0,0,1,1}(C_0, C', \{s, \neg s\})$ $-d_{0,0,1,1}(C_0, C, \{s, \neg s\})$
remove contradicting element from C	-1	+1
revise contradicting element in C	-1	+1
add missing element to C	-0.3	0
remove additional element from C	-1	+1

TABLE 8.1: Differences between contributions to Hamming distances from altered commitments C' and original commitments C. Negative numbers signify an improvement in the measure after the change, positive numbers indicate a worsening.

The complete linearity of the achievement function allows to distribute (push in) the weights α_A and α_F over the individual contributions of the hamming distances:

$$
\begin{aligned}
&Z(C, T|C_0) \\
&= \alpha_A \cdot A(C, T) + \alpha_F \cdot F(C|C_0) + \alpha_S \cdot S(T) \\
&= \alpha_A \cdot (1 - \frac{D_{0,0.3,1,1}(C, \overline{T})}{n}) + \alpha_F \cdot (1 - \frac{D_{0,0,1,1}(C_0, C)}{n}) + \alpha_S \cdot (1 - \frac{|T| - 1}{|\overline{T}|}) \\
&= \alpha_A - \frac{\alpha_A \cdot D_{0,0.3,1,1}(C, \overline{T})}{n} + \alpha_F - \frac{\alpha_F \cdot D_{0,0,1,1}(C_0, C)}{n} + \alpha_S - \frac{\alpha_S \cdot (|T| - 1)}{|\overline{T}|} \\
&= 1 - \frac{\alpha_A \cdot D_{0,0.3,1,1}(C, \overline{T}) + \alpha_F \cdot D_{0,0,1,1}(C_0, C)}{n} - \frac{\alpha_S \cdot (|T| - 1)}{|\overline{T}|}
\end{aligned}
$$

Changing the commitments has no effect on

$$\frac{\alpha_S \cdot (|T| - 1)}{|\overline{T}|},$$

and n is fixed. Consequently, Z can be optimised by changing the commitments such that the following term is minimised:

$$\begin{aligned}
&\alpha_A \cdot D_{0,0.3,1,1}(C, \overline{T}) + \alpha_F \cdot D_{0,0,1,1}(C_0, C) \\
&= \alpha_A \cdot \sum_{i=1}^{n} d_{0,0.3,1,1}(C, \overline{T}, \{s_i, \neg s_i\}) + \alpha_F \cdot \sum_{i=1}^{n} d_{0,0,1,1}(C_0, C, \{s_i, \neg s_i\}) \\
&= \sum_{i=1}^{n} \alpha_A \cdot d_{0,0.3,1,1}(C, \overline{T}, \{s_i, \neg s_i\}) + \alpha_F \cdot d_{0,0,1,1}(C_0, C, \{s_i, \neg s_i\})
\end{aligned}$$

We apply the weights to the contributions from Table 8.1 and arrive at Table 8.2. Since the achievement function is optimised for minimal contribution

change	**account** $d_{0,0.3,1,1}(C', \overline{T}, \{s, \neg s\})$ $-d_{0,0.3,1,1}(C, \overline{T}, \{s, \neg s\})$	**faithfulness** (worst case) $d_{0,0,1,1}(C_0, C', \{s, \neg s\})$ $-d_{0,0,1,1}(C_0, C, \{s, \neg s\})$
remove contradicting element from C	$-\alpha_A$	$+\alpha_F$
revise contradicting element in C	$-\alpha_A$	$+\alpha_F$
add missing element to C	$-0.3 \cdot \alpha_A$	0
remove additional element from C	$-\alpha_A$	$+\alpha_F$

TABLE 8.2: Weighted differences between contributions to Hamming distances from altered commitments C' and original commitments C.

and $\alpha_A > \alpha_F$, it is always more attractive to change the commitments to increase account rather than faithfully respecting the initial commitments. This argument can be repeated for every sentence for which C and $\overline{T}$ differ.

In summary, if (C, T) is a global optimum but $A(C, T) < 1$, then there is a position (C', T) such that $A(C, T) < A(C', T)$ contradicting (C, T) being a global optimum. Consequently, we must have $A(C, T) = 1$, i.e., T accounts fully and exclusively for S (FEA). This shows that (C, T) is a full RE state. □

Note that this argument does not work for quadratic model variants, and in particular, the default model of Beisbart, Betz, and Brun (2021). Remember that the Hamming distance D is a summation of penalties. Consequently, squaring the hamming distance yields a polynomial expression where every contributing penalty "interferes" by multiplication with the others. This blocks the above strategy of comparing the contributions and distributing the weights α_A or α_S over these expressions. Later, we can observe a gradual transition in the default model between configurations that yield global optima, which are almost certainly full RE states, to configurations that almost certainly fail in this respect.

Nonetheless, $\alpha_A = \alpha_S$ will be come relevant as a line of symmetry for the quadratic model in Section 8.2.

Pareto Efficiency and Global Optima I discussed Pareto efficiency as reasonable requirement for aggregation rules for ordering theories according to theoretical virtues in Chapter 6, and more generally, it is also proposed as a condition on equilibrium states by proponents on RE. The same applies to the formal model of RE, which measures epistemic desiderata on a ratio scale before aggregating them in the achievement function.

Before we proceed, we remind ourselves of the definition of Pareto efficiency and establish some useful notation. Assume that a dialectical structure τ and some initial commitments C_0 are given. Let (C, T) be an epistemic state. The values of account, systematicity and faithfulness (given the initial commitments) are fully determined, hence we can assign a column vector to (C, T) to store these values:

$$u_{(C,T)} = \begin{pmatrix} A(C,T) \\ S(T) \\ F(C \mid C_0) \end{pmatrix} \in \mathbb{R}^3$$

If the situation is unambiguous, the index of u is dropped, and we denote the individual components of u with u_A, u_S, and u_F, respectively. Note that multiple epistemic states can yield the same vector due to identical measures for account, systematicity and faithfulness. Analogously, we represent a configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$ as a vector

$$\alpha = \begin{pmatrix} \alpha_A \\ \alpha_S \\ \alpha_F \end{pmatrix} \in \mathbb{R}^3.$$

This allows to express the achievement function very concisely:

$$\alpha \cdot u = \alpha_A A(C,T) + \alpha_S S(T) + \alpha_F F(C \mid C_0) = Z(C, T \mid C_0),$$

where $\cdot$ denotes is the dot product.

Let $\mathcal{U}$ consist of $u_{(C,T)}$ for all admissible epistemic states (C, T) for a dialectical structure τ. An epistemic state (C, T) is admissible if and only if the commitments are minimally consistent and the theory is dialectically consistent. $u_{(C,T)}$ is *Pareto efficient* in $\mathcal{U}$ if and only if there is no $u_{(C',T')}$ in $\mathcal{U}$ for an

epistemic state (C', T'), such that

$$A(C', T') \geq A(C, T) \text{ and } S(T') \geq S(T) \text{ and } F(C' \mid C_0) \geq F(C \mid C_0),$$

where at least one inequality is strict. An other way to think of a Pareto efficient $u_{(C,T)}$ is the following: There is no epistemic state (C', T') such that switching from (C, T) to (C', T') would strictly improve account, systematicity or faithfulness without making any other of these measures worse.

It seems plausible to me, that an ideally rational agents should aim for epistemic states that exhibit Pareto efficient measures, especially if we are interested in whether a state of RE has been reached (see also condition 4 on page 22, which is endorsed by Elgin (1996, 2017) or Rechnitzer (2022)). If an epistemic state is not Pareto efficient, there is room for improvement, which provides incentive to revise an epistemic state disturbing its provisional state of equilibrium.

The following proposition establishes an interesting connection between global optima, Pareto efficient states and configuration of weights.

Proposition 5. *Assume that a dialectical structure τ and some initial commitments C_0 are given.*

i) If (C, T) is a global optimum according to the achievement function Z specified with a configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$ $(\alpha_i > 0, i \in \{A, S, F\})$, then $u_{(C,T)}$ is Pareto efficient in $\mathcal{U}$.

ii) Let $u_{(C,T)}$ be Pareto efficient in $\mathcal{U}$. Then there is a configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$ $(\alpha_i \geq 0, i \in \{A, S, F\})$ such that (C, T) is a global optimum according to the achievement function Z specified with $(\alpha_A, \alpha_S, \alpha_F)$.

Proof. i) Let (C, T) be a global optimum according to the achievement function Z specified with a configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$. For a proof by contradiction, assume that $u_{(C,T)}$ is not Pareto efficient in $\mathcal{U}$. This means that there is an epistemic state (C', T') such that

$$A(C', T') \geq A(C, T) \text{ and } S(T') \geq S(T) \text{ and } F(C' \mid C_0) \geq F(C \mid C_0),$$

where at least one inequality is strict. Without loss of generality, assume that $A(C', T') > A(C, T)$, which implies

$$Z(C', T' \mid C_0) = \alpha \cdot u_{(C',T')} > \alpha \cdot u_{(C,T)} = Z(C, T \mid C_0).$$

This contradicts (C, T) being a global optimum.

ii) Here, I can offer but a proof sketch because the result builds upon important insights from convex geometry such as Minkowski's hyperplane separation theorem, which I cannot present in full detail.[2]

Let $\mathcal{U}^*$ be the convex hull of $\mathcal{U}$, i.e., the set of all convex combinations of U. Formally,

$$\mathcal{U}^* = \{\sum_{i=1}^{n} \lambda_i \begin{pmatrix} u_A^i \\ u_S^i \\ u_F^i \end{pmatrix} \mid u^i \in \mathcal{U}, \lambda_i \in [0,1], \sum_{i=1}^{n} \lambda_i = 1\}.$$

$\mathcal{U}^*$ is convex, and the Pareto efficient elements in $\mathcal{U}$ are part of the boundary of $\mathcal{U}^*$ (see Figure 8.5). Let (C,T) be an epistemic state such that $u = u_{(C,T)}$ is a Pareto efficient in $\mathcal{U}$. Similar to $\mathcal{U}^*$, we define a set of elements in $\mathbb{R}^3$ that are component-wise greater or equal to u:

$$\mathcal{V} = \{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3 \mid x \geq u_A, y \geq u_S, z \geq u_F\}.$$

$\mathcal{V}$ is also convex, and $u \in \mathcal{U}^* \cap \mathcal{V}$. In fact, $\mathcal{U}^* \cap \mathcal{V} = \{u\}$ due to the Pareto efficiency of u. As the interior of $\mathcal{V}$ does not intersect $\mathcal{U}^*$, the hyperplane separation theorem applies: There is a vector $a \in \mathbb{R}^3$ that is the normal of a hyperplane separating $\mathcal{U}^*$ and $\mathcal{V}$ as illustrated in Figure 8.5.

Separation by a hyperplane means

$$a \cdot v \geq a \cdot u$$

for all $u \in \mathcal{U}^*$ and $v \in \mathcal{V}$. In fact, the components of a are greater or equal to 0 (not all being 0). By normalisation with the sum over components of a we can find another normal vector

$$\alpha = \begin{pmatrix} \alpha_A \\ \alpha_S \\ \alpha_F \end{pmatrix},$$

such that $\alpha_A \geq 0, \alpha_S \geq 0, \alpha_F \geq 0$ and $\alpha_A + \alpha_S + \alpha_F = 1$.

[2]The full proof originates from welfare economics and it can be found in (Negishi, 1960), who follows up on the famous theorems of (Arrow and Debreu, 1954) concerning the existence of an equilibrium for a competitive economy. For a more recent approach from convex optimisation, see (Boyd and Vandenberghe, 2004, 55–58, 177–179).

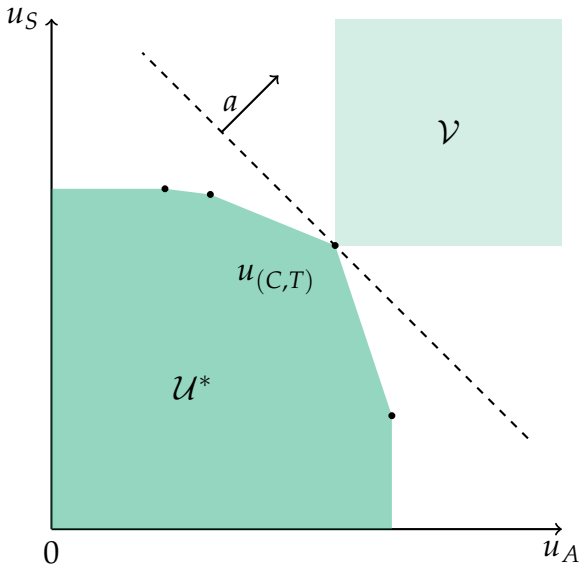


FIGURE 8.5: $\mathcal{U}^*$ and $\mathcal{V}$ touch in a Pareto efficient point $u_{(C,T)}$ and they can be separated by a hyperplane with a normal vector a. Note that the hyperplane is a line for two dimensions.

Consequently, the following holds for every epistemic state (C', T'):

$$u' = u_{(C',T')} \in \mathcal{U} \subseteq \mathcal{U}^*,$$

and thus,

$$\alpha \cdot u \geq \alpha \cdot u',$$

which is equivalent to

$$Z(C, T \mid C_0) \geq Z(C', T' \mid C_0).$$

This shows that (C, T) is a global optimum according to the achievement function Z specified with a configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$. □

What can we learn from these results? The configuration of weights give a general "direction" of inquiry for global optima. Geometrically, this means that the configuration, understood as a normal vector, determines a hyperplane. The hyperplane is translated (along the normal vector) to touch as few as possible points in the set of desideratum measures. The corresponding epistemic states are Pareto efficient, and they are global optima according to the achievement function specified with the configuration.

Configurations of weights handle trade-offs between epistemic states that exhibit Pareto efficient measures for the desiderata. Without such trade-offs there is no straightforward selection mechanism among Pareto optima. The

weights determine "exchange rates"(Boyd and Vandenberghe, 2004, 184) between desiderata. How much account are we willing to give up in order to increase systematicity by a specific amount? The configuration of weights encodes the answer.

The second part of Proposition 5 provides motivation to impose additional conditions on global optima to qualify as (full) RE state. There is some sort of "anything goes" with respect to Pareto efficient states. Every epistemic state that exhibits Pareto efficient measures for desiderata is a global optimum for a configuration of weights. By imposing additional constraints, e.g., (CCT) or (FEA) from Chapter 7) which are independent of the configuration of weights, "anything goes" is blocked for (full) RE states. This is helpful to narrow down a range of promising configurations for exploring simulations.

This result may be of importance to configuration of theoretical virtues more generally, e.g., for the evaluation of scientific theories. Trading-off gradual virtues by means of an additive, convex aggregation function allows for "anything goes" with respect to the weights. Every theory that exhibits Pareto efficient measures for gradual virtues is an optimum for a configuration of weights. This can be limited if we include additional, categorical virtues that serve as necessary requirements.

Finally, the result opens the door for a completely new line of research because it reveals a connection between the formal model and formal approaches to multicriteria optimisation in economics. For the latter it is a well-known and extensively studied idea that that every Pareto optimum solves a scalarised optimisation problem for some weights (Negishi, 1960), but the analogy goes deeper. Admissible epistemic states are similar to feasible allocations of goods, the desiderata can be seen as utilities that encode the preferences of consumers, and searching for a global optimum of the achievement function corresponds to maximising a social welfare function. The conditions of attaining equilibrium states in economies has been studied intensively, take for example the famous theorems of Arrow and Debreu (1954). After the discussion in Section 6.2 of Arrow's impossibility theorem (Arrow, 1951) transferred to theory choice by Okasha (2011), this is the second appearance of such formal results in this project. Hence, it would be interesting to see whether this analogy allows to transfer additional results to the discussion of RE. Unfortunately, this goes far beyond the scope of the current project. Nonetheless, Proposition 5 is another fruitful analytic finding about the formal model. It sheds some light on role of theoretical virtues in the formal

model. Proposition 5 relates desiderata, global optima and Pareto efficient states, and its proof gives a geometrical interpretation of global optimisation.

8.2 Robustness

Analysing a model's robustness is an important aspect of its validation. Robustness analysis is motivated by the worry that idealising and simplifying assumptions, which went into the construction of the model, may introduce artefacts of formalisation. This may cause a model to behave in certain ways by accident, rather than representing the essential features of the target system.[3] Moreover, robustness is important for prediction, or if we want to apply a model to new cases.

The formal model of RE is no exception in this regard as it relies on various idealising and simplifying assumptions. More specifically, a substantial amount of parameters is set at the outset including the penalties in the functions that measure account and faithfulness, the order of the function G, or the configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$ that guide the trade-offs between desiderata in the achievement function Z.

As the focus of this project rests on the role of theoretical virtues in RE, I will confine the subsequent robustness considerations to the most relevant parameters in this respect, i.e., configurations of weights.

What is the threat of a model that fails to be robust? In this case the behaviour of the model is extremely sensitive to small differences in weight configurations. Configurations that produce the same behaviour may happen to be "unconnected" rendering identical results a matter of coincidence. Moreover, in a model that is highly sensitive to weight configurations, salient differences in behaviour cannot be traced to, or explained by differences in weight configurations.

Convexity A set (in our case, points in three-dimensional space) is *convex* if and only if the line segment between any two points of the set lies entirely in the set, too. Figuratively, you cannot step outside of a convex set if you walk in a straight line from one point to another. Figure 8.6 gives a simple visual illustration of convex and non-convex sets.

The behaviour, which we are going to study with respect to weight configurations is the yielding of (sets of) global optima and equilibration fixed points. Assume that some set of initial commitments C_0 and a dialectical

[3] See, for example, (Levins, 1966), or (Weisberg, 2006).

FIGURE 8.6: The left shape is convex because the line segment between any two points of the set lies in the set, too. The right shape fails in this respect, and is non-convex because it has a "dent".

structure τ are given. Let us say that a configuration of weights *yields* a global optima (C, T) (or a fixed point), if (C, T) is a global optima according to the achievement function Z specified by the configuration of weights (results from a equilibration process with said achievement function). Note that a configuration of weights may yield multiple global optima or fixed points due to ties in the achievement function or random choices during equilibration.

Assume that a set of configuration exhibits the same behaviour. Why is convexity a desirable property from the viewpoint of model robustness with respect to weight configurations? First, a convex set is connected in the mathematical sense. In the space of weight configurations, this means that a convex set of weight configurations does not fall apart into separated regions that yield the same behaviour. In fact, convexity implies even more: every point in a convex set is visible or reachable in a straight line segment from any other point in the set. Moreover, a convex set of weight configurations that yield the same behaviour does not have "holes" for which the model behaves differently. This all speaks in favour of the model being robust with respect to all configurations of a convex set. There, its behaviour is insensitive to arbitrarily small changes in the weights.

Concerning the aspect of prediction in model robustness, (Beisbart, Betz, and Brun, 2021) present an analytical result:

Proposition 6 (Beisbart, Betz, and Brun, 2021, 468–470)**.** *If each combination of weights from*

$$\{(\alpha_A^i, \alpha_S^i, \alpha_F^i) \mid i = 1, \ldots, n\} \quad (n \in \mathbb{N})$$

yields the same set of global optima (the same unique equilibration process with no random choices) for a fixed dialectical structure and fixed initial commitments, then every combination of weights in the convex hull of the set $\{(\alpha_A^i, \alpha_S^i, \alpha_F^i) \mid i =$

$1, \ldots, n\}$ *yields the same set of global optima (resp. the same equilibration process), too.*

Let V be the set of weight configurations mentioned in the proposition above. What is the convex hull of V? There are multiple equivalent definitions, but here is one that is useful to see how we can construct new configurations. The convex hull of V is the set of all convex combinations of V. Formally, the set of all convex combinations is given by

$$\{\sum_{i=1}^{n} \lambda_i \begin{pmatrix} \alpha_A^i \\ \alpha_S^i \\ \alpha_F^i \end{pmatrix} \mid \begin{pmatrix} \alpha_A^i \\ \alpha_S^i \\ \alpha_F^i \end{pmatrix} \in V, \lambda_i \in [0,1], \sum_{i=1}^{n} \lambda_i = 1\}.$$

Slightly less technical, for two configurations, the set of all their convex combinations (and hence, their convex hull) is given by the line segment that connects them. We can express the convex combinations of two points formally by

$$\begin{pmatrix} \alpha_A^\lambda \\ \alpha_S^\lambda \\ \alpha_F^\lambda \end{pmatrix} = (1-\lambda) \cdot \begin{pmatrix} \alpha_A^1 \\ \alpha_S^1 \\ \alpha_F^1 \end{pmatrix} + \lambda \cdot \begin{pmatrix} \alpha_A^2 \\ \alpha_S^2 \\ \alpha_F^2 \end{pmatrix}$$

where $\lambda \in [0,1]$. The convex hull of some configurations is illustrated in Figure 8.7.

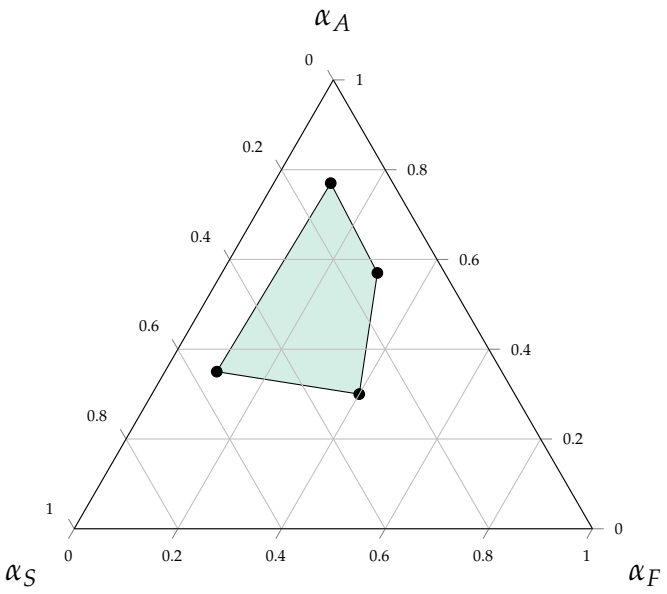


FIGURE 8.7: Illustrating the proposition of (Beisbart, Betz, and Brun, 2021). If some configurations (black dots) yield the same set of global optima (the same equilibration process without random choices), then any configuration on the inside (green region) will exhibit the same behaviour.

We can use this proposition to predict how the model behaves if we know some configurations that yield the same set of global optima or the same equilibration process without random choices. In this case, any convex combination of the known configurations will exhibit the same behaviour. Moreover, the result has important practical consequences when we study the model by computer simulations. Theoretically, there are uncountably many configurations of weights (due to working with the real numbers), but we can only simulate RE with a finite number of weight configurations. At best we can generate a discrete and uniform distribution of weight configurations according to a fixed resolution, but there will always be unsimulated configurations between two data points. The proposition helps to fill these gaps, if the simulated data points exhibit the same behaviour.

The proposition leaves open whether a global optimum can be reached from different configurations of weights that do not yield the same set of global optima. In this case, it would still be detrimental to the model's robustness if a global optimum could be reached from configurations that do not hang together as Proposition 6 applies to regions of configurations that yield the same global optima.

Here, I like to complement the proposition of Beisbart, Betz, and Brun (2021) with a more general analytical finding concerning global optima, that reverses the direction in some sense. The proposition discussed so far, departs from a set of configurations that yield the same set of global optima and tells us something about the set of global optima from convex combinations of those configurations. Here, we start from a single global optimum and ask what we can learn about configurations that yield this optimum. Given that an epistemic state is a global optimum for a configuration of weights, what can we say about the set of configurations for which the epistemic state is also a global optimum?

Proposition 7. *Assume that a dialectical structure τ and some initial commitments C_0 are given and fixed. The set of configurations of weights, for which an epistemic state (C, T) is among the global optima according to the achievement function Z (specified for every such configuration), is convex.*

Proof. Let V be the set of all weight configurations that yield (C, T) as a global optimum.

Case 1: There is only one configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$ for which (C, T) is a global optimum according to the achievement function Z (specified with $(\alpha_A, \alpha_S, \alpha_F)$). In this case, the set $V = \{(\alpha_A, \alpha_S, \alpha_F)\}$ is convex by definition.

Case 2: V contains more than one element such that (C, T) is among the global optima according to the achievement function specified for the configuration. Let $(\alpha_A^1, \alpha_S^1, \alpha_F^1)$ and $(\alpha_A^2, \alpha_S^2, \alpha_F^2)$ be two arbitrary, distinct configurations of weights from V. Consider the their set of convex combinations:

$$K = \{(\alpha_A^\lambda, \alpha_S^\lambda, \alpha_F^\lambda) \mid \lambda \cdot (\alpha_A^1, \alpha_S^1, \alpha_F^1) + (1 - \lambda) \cdot (\alpha_A^2, \alpha_S^2, \alpha_F^2), \lambda \in [0, 1]\}$$

If $(\alpha_A^1, \alpha_S^1, \alpha_F^1)$ and $(\alpha_A^2, \alpha_S^2, \alpha_F^2)$ yield the same set of global optima (including (C, T)), then so will every configuration from K (above proposition of (Beisbart, Betz, and Brun, 2021, 468)). Otherwise $(\alpha_A^1, \alpha_S^1, \alpha_F^1)$ and $(\alpha_A^2, \alpha_S^2, \alpha_F^2)$ yield different sets of global optima (both including (C, T)). Let $(\alpha_A^\lambda, \alpha_S^\lambda, \alpha_F^\lambda) \in K$ for some $\lambda \in (0, 1)$ (the cases $\lambda = 0, 1$ are trivial) and assume for a contradiction that (C, T) is *not* a global optimum according to the achievement function specified by $(\alpha_A^\lambda, \alpha_S^\lambda, \alpha_F^\lambda)$. Lets denote the achievement functions that are specified by $(\alpha_A^1, \alpha_S^1, \alpha_F^1)$, $(\alpha_A^2, \alpha_S^2, \alpha_F^2)$, and $(\alpha_A^\lambda, \alpha_S^\lambda, \alpha_F^\lambda)$ with Z^1, Z^2 and Z^λ, respectively. Note that the following holds (Beisbart, Betz, and Brun, 2021, 469):

$$Z^\lambda = \lambda \cdot Z^1 + (1 - \lambda) \cdot Z^2 \tag{8.3}$$

If (C, T) is not a global optimum according to Z^λ, then there is an epistemic state (C', T') such that

$$Z^\lambda(C, T \mid C_0) < Z^\lambda(C', T' \mid C_0).$$

By Equation 8.3, this is equivalent to

$$\lambda \cdot Z^1(C, T \mid C_0) + (1 - \lambda) \cdot Z^2(C, T \mid C_0) \\ < \lambda \cdot Z^1(C', T' \mid C_0) + (1 - \lambda) \cdot Z^2(C', T' \mid C_0).$$

From the fact that all individual contributions to above inequality are positive, we can derive

$$Z^1(C, T \mid C_0) < Z^1(C', T' \mid C_0)$$

or

$$Z^2(C, T \mid C_0) < Z^2(C', T' \mid C_0)$$

In either case we have a contradiction to (C, T) being a global optimum according to Z^1 or Z^2. Consequently, (C, T) is a global optimum according to Z^λ for every configuration in K, the convex combinations of $(\alpha_A^1, \alpha_S^1, \alpha_F^1)$ and $(\alpha_A^2, \alpha_S^2, \alpha_F^2)$. As these two configurations are arbitrary in V, we have shown

that (C, T) is among the global optima for any convex combination of configurations in V, which is to say that V is convex. □

Proposition 7 immediately yields interesting corollaries: First, it generalises the part about global optima in Proposition 6 of (Beisbart, Betz, and Brun, 2021). Start with a set of two or more global optima. Each global optimum stems from a convex set of configurations. It is a basic result of convex geometry, that the intersection of a collection of convex sets is again convex. Thus, the intersection of sets of configurations is convex. If it is non-empty, its configurations yield all global optima from the set we started with.[4]

Next, the fact that there are finitely many epistemic states (of which global optima are a subset) implies that the convex sets of configurations for global optima cannot all be singletons. Hence, there are convex sets of configurations that are extended having a positive "width" that "cover" the space of configurations. A convex set of configuration robustly yields the global optima, and its "degree" of robustness can be quantified.

Finally, the dimension of the space of weight configurations (in our case: 3) does not matter. Consequently, the proposition generalises to extensions of the model with additional weights for new measures of desiderata as long as they are aggregated in the achievement function as a convex combination.

Symmetry Consider to configuration of weights,

$$(\alpha_A, \alpha_S, \alpha_F) \text{ and } (\alpha_F, \alpha_S, \alpha_A),$$

which swap the weights for account and faithfulness. Geometrically, this corresponds to a reflection on the line defined by $\alpha_A = \alpha_F$ in a ternary plot.

α_A and α_F determine the trade-off between account and faithfulness in a global optimum (for a fixed weight α_S for systematicity). Reflection on the line $\alpha_A = \alpha_F$ reverses the roles of α_A and α_F, as well as the direction of the trade-off.

It turns out that we can relate the global optima of such "reflected" configurations. In particular, they can be constructed from each other. In ternary plots, this results in a symmetry of shape of convex sets of configurations that yield related global optima (for visualisations, see Section 8.3).

A first step towards an explanation of this symmetry is the following proposition.

[4]The empty set is convex by definition. This covers the uninteresting case of non-intersecting sets of configurations.

Proposition 8. *Assume that a dialectical structure τ and some initial commitments C_0 are given. Let (C, T) be a global optimum according to the achievement function Z (specified for some configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$). Then the dialectical closure $\overline{T}$ does not expand C (meaning that there are no penalties for expansions in $A(C, T)$.*

Proof. For a proof by contradiction, assume that $\overline{T}$ expands C with respect to at least one sentence t, that is $t \in \overline{T}$ and $t \notin C$, resulting in a penalty for expansion ($d_1 = 0.3$ in the default model in $A(C, T)$.

Consider $C' = C \cup \{t\}$. C' is minimally consistent. Otherwise, $\neg t \in C$, but then T would not extend but contradict C with respect to t, resulting in a penalty for contradiction in $A(C, T)$. Then the following hold:

i) $A(C', T) > A(C, T)$ as C' remedies that T extends C with respect to t.

ii) $S(T)$ is unaffected by extending the commitments.

iii) $F(C' \mid C_0) = F(C \mid C_0)$

This is the reasoning for iii): In the default model, expansions are not penalised in F ($d_1 = 0$) and new contractions cannot result from adding elements to the commitments. This leaves contradictions: Assume that adding t to C causes a contradiction with C_0. Hence, $\neg t \in C_0$, but also $\neg t \notin C$. This means that C contracts C_0 with respect to $\neg t$. As the penalties for contradictions and contractions are identical ($d_3 = d_2 = 1$), the contradiction between C' and C_0 is compensated by the contraction between C and C_0. Thus, we have $F(C' \mid C_0) = F(C \mid C_0)$.

Finally, i) – iii) jointly imply $Z(C', T \mid C_0) > Z(C, T \mid C_0)$, which contradicts (C, T) being a global optimum. □

Note that the Proposition 8 is independent of specific configuration of weights. This is due to the fact that expansion penalties for account ($d_1 = 0.3$) and faithfulness ($d_1 = 0$) differ categorically. As the expansion penalty for faithfulness is zero, no difference between α_F and α_A can make it more attractive to accept an expansion penalty in account. The situation would be fundamentally different, if the expansion penalty for faithfulness was non-zero.

The proposition does not show that the expansion penalty for account is useless. It separates sub-optimal states from global optima as a necessary condition, it is at work during equilibration processes, and occasionally occurs in sub-optimal fixed points.

As a consequence of Proposition 8, the penalty $d_1 = 0.3$ in the measure for account $A(C, T)$ does not matter for global optima. Hence, the penalties for account $A(C, T)$ and faithfulness $F(C \mid C_0)$ are effectively identical ($d_0 = d_1 = 0$ and $d_2 = d_3 = 1$) for global optima. This provides the ground for symmetry.

Assume that an agent is in a situation where they have some initial commitments C_0 and a non-empty theory T. How could they form a set of commitments C in an attempt to optimise the achievement function Z for some configuration $(\alpha_A, \alpha_S, \alpha_F)$? Let us illustrate the situation with the standard example (see Figure 7.1), case C (see Table 7.1). The initial commitments are $C_0 = \{3, 4, 5, 6, 7\}$ and let the theory be $T = \{1\}$ (which is part of the global optimum for the standard configuration). Figure 8.8 depicts the initial commitments C_0, the theory T, and a new basis C^* for the commitments. C^* contains the sentences on which C_0 and $\overline{T}$ agree, as well as the sentences for which $\overline{T}$ expands C_0.

$$C_0 = \{3, 4, 5, 6, 7\} \longrightarrow C^* = \{1, -2, 3, 4, 5\} \longleftarrow \{1, -2, 3, 4, 5, -6\} = \overline{T}$$

FIGURE 8.8: C^* includes agreements between C_0 and $\overline{T}$ (i.e., 3, 4, and 5), as well as expansions from C_0 by $\overline{T}$ (i.e., 1 and -2).

C_0 and $\overline{T}$ disagree about the remaining sentences 6 and 7. With respect to 6, C_0 and $\overline{T}$ contradict each other, and concerning 7, $\overline{T}$ contracts C_0. What are the options to modify C^* such that the modifications are optimal according to the achievement function? For every sentence s for which C_0 is contracted by $\overline{T}$ (i.e., $s \in C_0$, $\{s, \neg s\} \cap \overline{T} = \emptyset$), we can

(i) not add s to C^*, or

(ii) add s to C^*.

For every sentence s, for which C_0 and $\overline{T}$ contradict each other (without loss of generality, $s \in C_0$, $\neg s \in \overline{T}$), we can

(iii) add $\neg s$ to C^*

(iv) add s to C^*.

These modifications affect the achievement function in different ways as penalties for account or faithfulness are distributed differently. Let A_- and $A_\lightning$ denote the contributions to the penalties in the measure for account for contractions (-) and contradictions ($\lightning$). Analogously, F_- and $F_\lightning$ denote the penalty contributions for faithfulness.

(i) $A_{\text{-}} : +0, \quad F_{\text{-}} : +1$

(ii) $A_{\text{-}} : +1, \quad F_{\text{-}} : +0$

(iii) $A_{\lightning} : +0, \quad F_{\lightning} : +1$

(iv) $A_{\lightning} : +1, \quad F_{\lightning} : +0$

Penalties are detrimental to the respective measure, and hence, (i) and (iii) reduce faithfulness in favour of account, (ii) and (iv) give up account in favour of faithfulness. Note that other modifications are not optimal according to the achievement function, as they amass more penalties. In the case of a contradiction, for example, neither adding s nor $\neg s$ to C^* would cause a contraction penalty for faithfulness ($F_{\lightning} : +1$) and an expansion penalty for account ($A_{+} : +0.3$).

Whether (and to which extent) the penalties, are distributed in favour of account or faithfulness depends on the specific configuration of weights (and their ratio). If, for example, account receives relatively more weight than faithfulness, modifications (i) and (iii), which avoid penalties for account, are relatively preferred over (ii) and (iv). In the standard example with the standard configuration ($\alpha_A = 0.35$, $\alpha_F = 0.10$), account is preferred for the adjustment of C^*. The results of these adjustments are depicted in figure 8.9. What happens if we reverse the modifications of C^*, i.e., if we replace (i) by

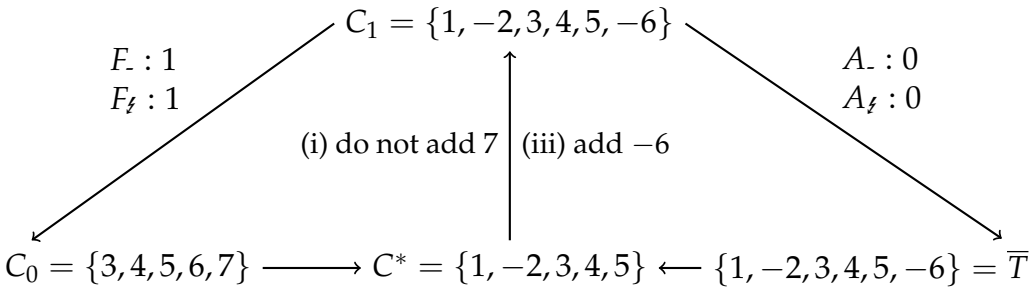


FIGURE 8.9: C_1 incorporates modifications of C^* that favour account and relegates penalties to faithfulness.

(ii), and (iii) by (iv), and vice versa? The penalties are transferred from faithfulness to account and vice versa. Figure 8.10 illustrates such a modification of C^* in our example. Interestingly, the formal model reaches C_2 (in global optima or fixed points) if the weights for account and faithfulness are interchanged, i.e. $\alpha_A = 0.10$ and $\alpha_F = 0.35$. Now, we can start to see from where the "symmetry" arises. Note that reversing the modifications flips the distribution of penalties in figure 8.10. Moreover, reversing the modifications is a self-inverse transformation.

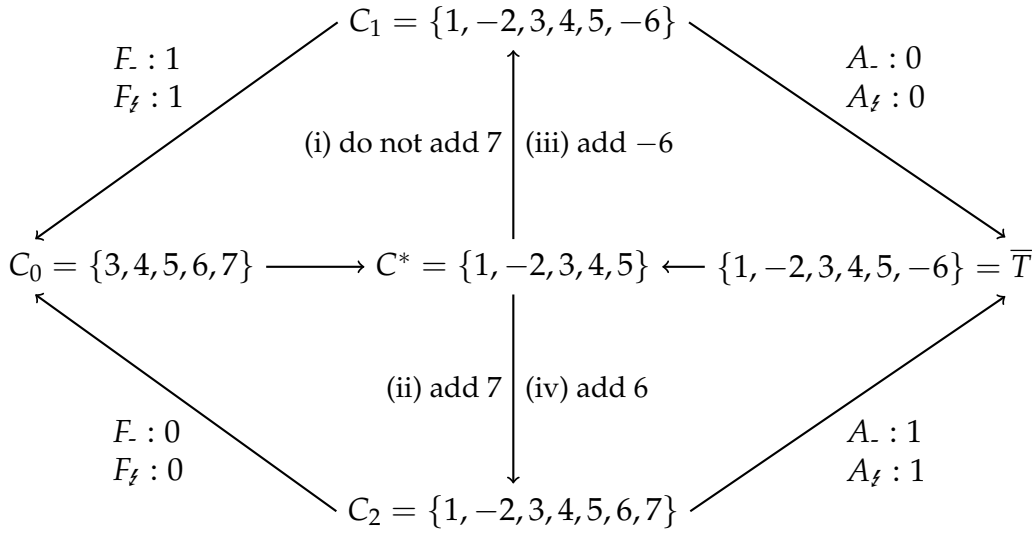


FIGURE 8.10: C_2 results from reversing the modifications that lead to C_1. C_2 favours faithfulness and relegates the penalties to account.

Proposition 9. *Assume that a dialectical structure τ and initial commitments C_0 are given. Let (C, T) be a global optimum (relative to C_0) according to an achievement function Z specified with $(\alpha_A, \alpha_S, \alpha_F)$ such that $A(C, T) < 1$ or $F(C \mid C_0) < 1$. Then, there are commitments C' such that (C', T) is a global optimum according to an achievement function Z' specified with $(\alpha_F, \alpha_S, \alpha_A)$.*

Proof. We start with the constructive method described above to form a basis C^* for commitments that consists of agreements between C_0 and $\overline{T}$ as well as expansions of C_0 by $\overline{T}$. As $A(C, T) < 1$ or $F(C \mid C_0) < 1$ the modifications that yield C from C^* involve the distribution of some penalties in favour of account or faithfulness. Let C' be the commitments that arise from C^* by reversing the modifications that lead from C^* to C. C' is uniquely determined.

Next, we prove that the achievement functions with interchanged weights yield identical values for the reversely modified commitments,

$$Z(C, T \mid C_0) = Z'(C', T \mid C_0) \tag{Z}$$

To establish this result, note that $S(T)$ is identical in Z and Z'. $A(C, T)$ does not involve expansion penalties (Proposition 8), and neither does $A(C', T)$ as it is constructed additively on top of C^*. This means that the measures for account and faithfulness effectively involve the same Hamming distance with penalties (+1) for contraction and contradictions. Reversing the modifications between C and C' shifts the account penalties for contradictions and contractions in $A(C, T)$ to the faithfulness penalties in $F(C' \mid C_0)$ (see, for

example, Figure 8.10). As there are no other penalties

$$A(C, T) = F(C' \mid C_0)$$

follows. Analogously, we have

$$F(C \mid C_0) = A(C', T).$$

This suffices to establish (Z):

$$\begin{aligned} Z(C, T \mid C_0) &= \alpha_A \cdot A(C, T) + \alpha_S \cdot S(T) + \alpha_F \cdot F(C \mid C_0) \\ &= \alpha_A \cdot F(C' \mid C_0) + \alpha_S \cdot S(T) + \alpha_F \cdot A(C', T) \\ &= Z'(C', T \mid C_0) \end{aligned}$$

It remains to show that (C', T) is a global optimum according to Z'. For a proof by contradiction, assume that (C', T) is not a global optimum according to Z' specified by $(\alpha_F, \alpha_S, \alpha_A)$. In this case, there is an epistemic state (C'', T'') such that $Z'(C'', T'' \mid C_0) > Z(C', T \mid C_0)$. By repeating the constructive method from above, we can find commitments C''' yet again, such that for the original achievement function Z the following holds:

$$\begin{aligned} Z(C''', T'' \mid C_0) &\overset{(Z)}{=} Z'(C'', T'' \mid C_0) \\ &> Z'(C', T \mid C_0) \\ &\overset{(Z)}{=} Z(C, T \mid C_0) \end{aligned}$$

This is a contradiction to (C, T) being a global optimum according to Z. Consequently, (C', T) is a global optimum according to Z'. □

Proposition 9 may seem rather unsettling from the perspective of model robustness. First, we can construct a globally optimal epistemic state with reversed contraction and contradiction penalties for every global optima that involve trade-offs between account and faithfulness. In this case, the global optima are guaranteed to disagree about some commitments. Next, the corresponding achievement functions do not help to decide between these optima because they yield the same value.

As an illustrative example, take again the standard example from above, case C, with $C_0 = \{3, 4, 5, 6, 7\}$ and $T = \{1\}$. $C_1 = \{1, -2, 3, 4, 5, -6\}$ is a global optimum according to the achievement function Z specified by the

standard configuration of weights $(\alpha_A, \alpha_S, \alpha_F) = (0.35, 0.55, 0.10)$. Furthermore, (C_1, T) is a full RE state because are C_1 and T are consistent with each other, and T fully accounts for C_1. $C_2 = \{1, -2, 3, 4, 5, 6, 7\}$ performs equally well according to the achievement function if it specified with reversed weights for account and faithfulness $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.55, 0.55)$. (C_2, T) is a global optimum, and note that both global optima reach a value of 0.9918 if the achievement function is specified accordingly. However, C_2 is dialectically inconsistent (1 and 6 cause a contradiction), it is inconsistent with the theory T (which implies -6), and it is not fully accounted for by T (7 is not part of $\overline{T}$). Compared to the initial commitments C_0 which are consistent, it is fair to say that C_2 does not achieve epistemic progress but a worsening.

Thus, value of the achievement function is, in general, not a viable measure for coherence or justification. Especially, across different configurations of weights, the values of achievement functions do not tell us whether one epistemic state is epistemically more desirable than an other from the viewpoint of RE. For some configurations the achievement function is not able to remove inconsistencies and misfits. This underlines the importance of additional requirements for RE states that have been proposed by Beisbart, Betz, and Brun (2021), e.g., (CCT) or (FEA). Hence, I will rely on additional consideration in order to identify prospective configurations of weights for further ensemble studies.

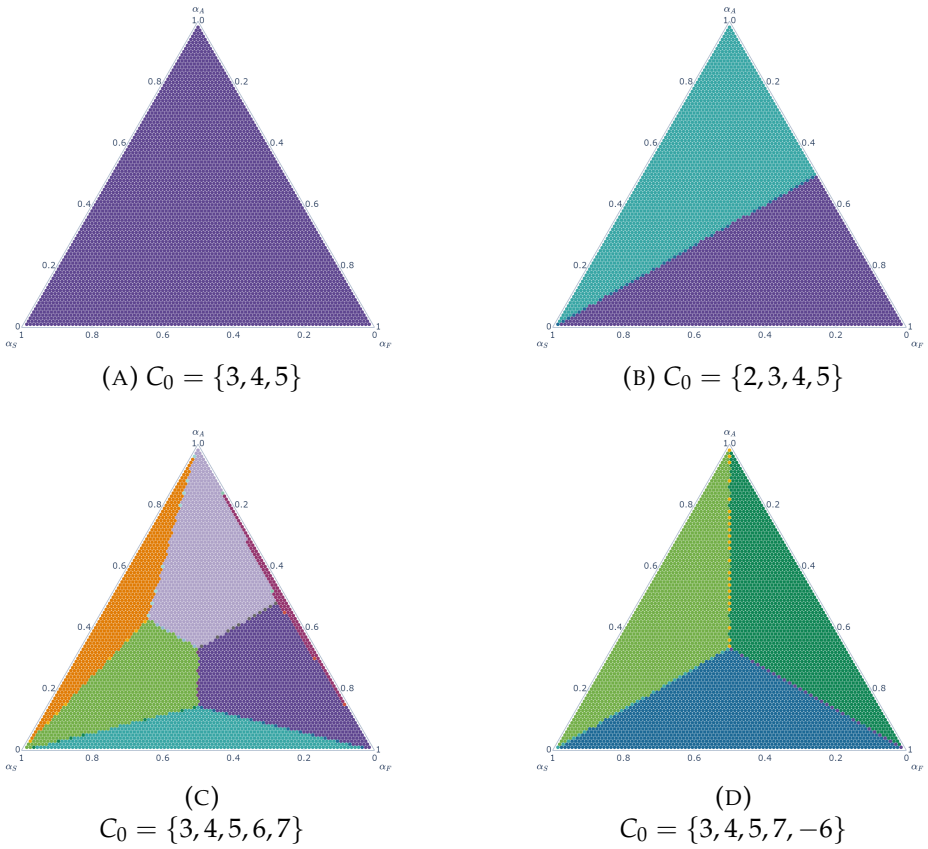
8.3 Illustrations

I conclude this chapter with a series of plots to illustrate the rather technical results of this chapter. In particular, convexity and symmetry are geometric properties that allow for insightful visualisations.

Convexity and Symmetry Without going too much into the details of producing data with the computer implementation (which is subject of Chapter 9), I report results from the following ensemble of simulations. For each of the four sets of initial commitments in the dialectical structure of the standard example, we determine all global optima and equilibration fixed points with a fairly high resolution of $\frac{1}{100}$ for the configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$, i.e., $(0.01, 0.01, 0.98)$, $(0.02, 0.01, 0.97)$, and so on. The extreme values 0.0 and 1.0 for weights have been excluded. This amounts to $\frac{1}{2} \cdot (99 \cdot 98) = 4851$

different weightings for each set of initial commitments, resulting in 19404 simulation setups.

In Figure 8.11, the configuration of weights are grouped by colour if they yield the same set of global optima. It is easy to see that they partition the space of weight configurations into convex regions (note that lines are convex, too) if we excuse distorting effects of floating point arithmetic on computers or the fact that individual markers are discretely spaced circles. This is in accordance with the Proposition 6 of (Beisbart, Betz, and Brun, 2021).



(A) $C_0 = \{3, 4, 5\}$ (B) $C_0 = \{2, 3, 4, 5\}$

(C) $C_0 = \{3, 4, 5, 6, 7\}$ (D) $C_0 = \{3, 4, 5, 7, -6\}$

FIGURE 8.11: Convex regions of weight configurations that yield the same set of global optima for four different initial commitments in the standard example. Note that colours do not denote the same set of global optima across different subfigures.

Case (A) in Figure 8.11 illustrates the special case of a "perfect" epistemic state that completely maximises the achievement function Z with a value of 1.0. In this case, there is exactly one set of global optima (that are also a full RE states), and it is yielded independent of specific configurations of weights as established in Proposition 2.

We also have an illustration of Proposition 9. The plots are saliently symmetrical with respect to the line $\alpha_A = \alpha_F$, which originates from the bottom

left corner, and which is perpendicular to the opposite side. Proposition 9 explains the symmetry observed for convex regions of configurations. For every global optimum according to an achievement function specified with a configuration of weights there is another global optimum according to an achievement function for interchanged weights for account and faithfulness. Interchanging these weight corresponds to a reflection on the $\alpha_A = \alpha_F$ line in the ternary plots. The shapes of regions that yield the same set of global optima are "reflected" with respect to this line.

For example, in case (B) of Figure 8.11, the lighter, turquoise region corresponds to the global optimum

$$(C, T) = (\{1, 3, 4, 5, -6, -2\}, \{1\}),$$

while the darker, purple region comprises weightings that yield

$$(C', T) = (\{1, 2, 3, 4, 5, -6\}, \{1\})$$

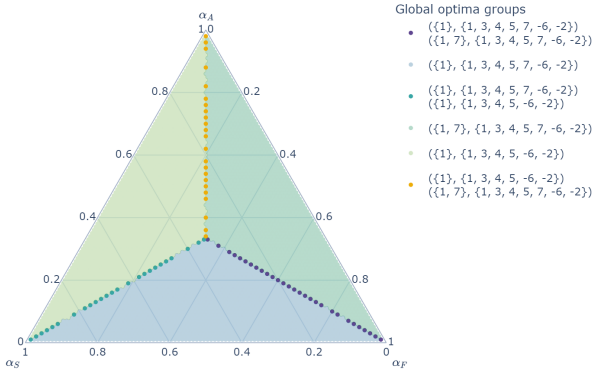
as global optimum. C revises 2, which is part of the initial commitments, to increase account (-2 is a consequence of T). In contrast, C' sticks to the initial commitment, despite adopting T, which results in a global optimum that is not an (full) RE state, e.g., due to the dialectical inconsistency of 2 and 4 in C'.

There are also open questions. Some regions include part of the symmetry line ($\alpha_A = \alpha_F$), e.g., the light green region in Figure 8.11, (C). In this case, the region is genuinely mirror symmetrical, as it is reflected into itself. Still, there are global optima with reversed penalty distributions in the same region. In what respects do such single regions differ from pairs of regions that are spatially separated, e.g., the orange and turquoise region in Figure 8.11, (C)? I conjecture that reflection pairs of regions distribute all penalties to either faithfulness or account in an all-or-nothing manner. In contrast, I suppose that single regions that contain part of the symmetry line distribute the penalties to both faithfulness and account.

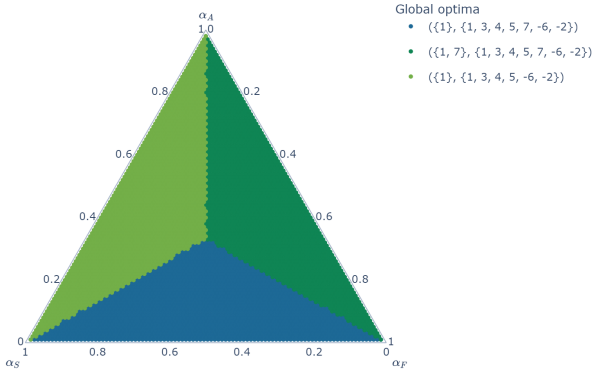
Figure 8.12 illustrates Proposition 7. Convex sets of weightings that yield a specific global optimum jointly cover the entire parameter space.

Note that there are "lines" of highly specific configurations between adjacent convex regions of configurations that yield the same set of global optima in Figure 8.12, (A). We are in a position to explain this by going back to the

[5]Note that the lines are not continuous due to the resolution of weight configurations. Small errors in floating point representation of real numbers causes the apparent gaps.



(A)



(B)

FIGURE 8.12: Standard example, case D. (A) Convex regions of weight configurations that yield the same global optima. The highlighted line regions consist of global optima from adjacent regions.[5] (B) Convex regions of weight configurations corresponding individual global optima. The lines from (A) disappear as the regions overlap.

proof of Proposition 5 that relates global optima and Pareto efficient states. There, Figure 8.5 indicates that there may be some leeway for a separating hyperplane to pivot around a Pareto efficient point. Every normal vector of a separating hyperplane translates into a configuration of weights. This results in the fact that multiple configuration of weights result in the same set of global optima. For specific configurations of weights the corresponding hyperplane touches multiple Pareto efficient points. Figure 8.13 depicts such a case in two dimensions, for which the hyperplane (a line) is uniquely determined by two points. In the case of the formal model's three dimensions, the hyperplane may be able to pivot around a line between two Pareto efficient points. This results in a line of configuration in a ternary plot that yield the global optima of adjacent regions.

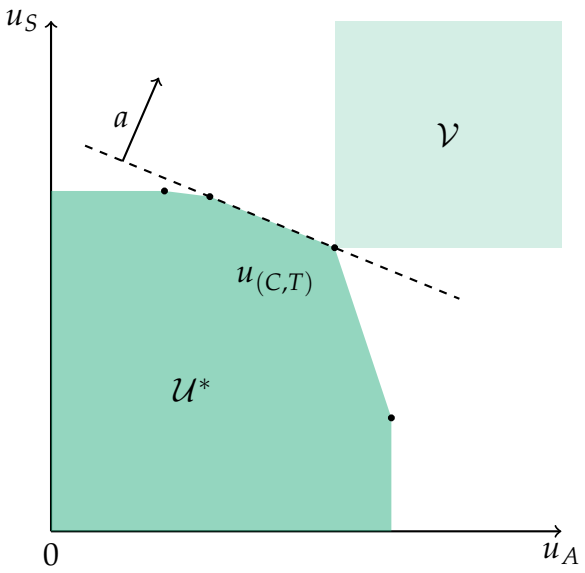
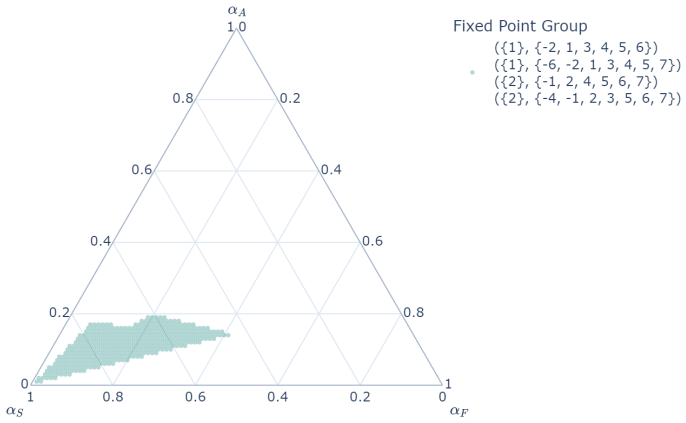
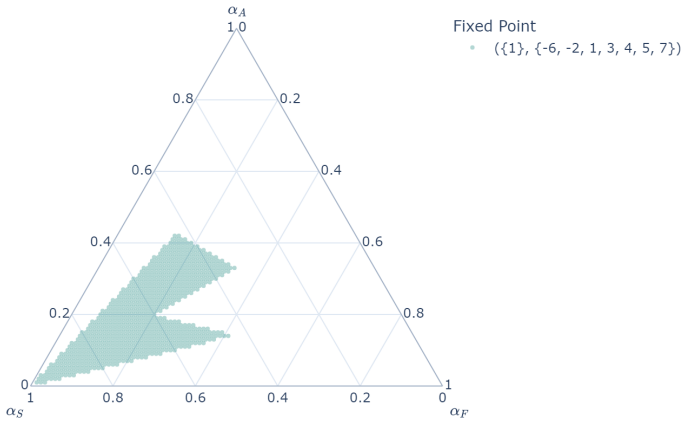


FIGURE 8.13: For specific configurations (translating to a normal vector a), the separating hyperplane goes through multiple Pareto efficient points. All epistemic states that yield these Pareto efficient points are global optima according to achievement function specified with this configuration.

Fixed points Concerning fixed points of equilibration processes, the discussed proposition of (Beisbart, Betz, and Brun, 2021) involves the additional requirement that no random choices occur during the process. Otherwise the proposition does not hold as illustrated in Figure 8.14 for the standard example, Case (C), with high values for α_S. Moreover, this counterexample establishes the same negative result for individual fixed points. This means that the new proposition does not hold for fixed points.



(A)



(B)

FIGURE 8.14: In the standard example, Case (C), the set of weight configurations that yield the above set of fixed points (due to random choices during RE processes) is not convex (A), and neither is the set of configurations that yield the individual fixed point $(\{1\}, \{-6, -2, 1, 3, 4, 5, 7\})$ (B).

Part III

Exploration

Chapter 9

Preparing the Formal Model for Simulations

The formal model of RE allows to be implemented as a computer program that is able to run simulations. But further preparations are needed before we turn to addressing objections to RE on the basis of simulations. There are many parameters for simulations that need to be determined besides providing dialectical structures and initial commitments. Of particular interest is the configuration of weights for desiderata in the aggregating achievement function. As the weights are real-valued numbers, there are uncountably many configurations of weights. Are there plausible ranges of configurations of weights for which the model performs well? And how do we asses its performance?

The aim of this chapter is to select promising configurations of weights in view of equilibration processes yielding full RE states.

I organise the chapter as follows: In Section 9.1, I shortly introduce the computer implementation of the formal model. I illustrate that the program is operative by replicating the simulation results in Section 9.2. In Section 9.3, I select promising configuration of weights by analysing results of simulations across the entire parameter space.

Here is an important preliminary remark: It is in the nature of formal models that they quickly entice to add extensions or variations once they are implemented to run on computers. Introducing weights for commitments and additional virtues, or splitting systematicity into separate measures for simplicity and scope are all within reach by changing a few lines of code.

Nonetheless, I will work with the formal model of Beisbart, Betz, and Brun (2021) as it stands, even in view of its known shortcomings, e.g., that the measure for systematicity does not discriminate singleton theories according to scope (see Section 8.1). It seems clear to me that there will be subsequent model variations that improve on the "default" model, but this requires a

solid baseline in the first place. It is of little use to fix one especially apparent shortcoming, only to find later that the fixing causes new issues. Aiming to understand some aspects of RE better by means of exploring the formal model, I prefer simplicity (as a virtue of a model) over premature complications. Moreover, I suppose that the unaltered formal model already performs quite well.[1]

9.1 The Python Implementation

The computer implementation that accompanies the model of Beisbart, Betz, and Brun (2021) is written for *Wolfram Mathematica*, which is a piece of proprietary software.[2] The members of the project "How far does reflective equilibrium take us? Investigating the power of a philosophical method" including myself decided to implement the formal model in Python, which is currently a very popular and freely available programming language.[3] As a general-purpose programming language, its uses range from scripting to scientific computing. Python code is (or, at least, should be) friendly to the reader, and the language is supposed to be easy to pick up, even for beginners.

It is a welcome upshot of the present implementation of the formal model of RE in Python and its well maintained documentation (thanks to Sebastian Cacean!) that people, which were not directly involved in developing the packages, were able to conduct research on their own. This illustrates that the computer implementation can serve as a publicly available tool for interested researches to explore the formal model of RE.

The packages were developed under the paradigm of object-orientation, that is, components of the formal model such as positions or dialectical structures were devised as classes that have specific properties (e.g., the size of the sentence pool), and methods (e.g., for returning the dialectical closure of a

[1] In addition, a ensemble study dedicated to comparing the default model to linear and quadratic model variants with measures of systematicity, which overcome the shortcoming concerning singleton theories, did not yield results that would recommend switching from the default model without further considerations a report can be found at https://www.philosophie.unibe.ch/unibe/portal/fak_historisch/dkk/philosophie/content/e40373/e82357/e776174/e1164904/e1365144/AssessingaFormalModelofReflectiveEquilibrium_ger.pdf. This would require a more in-depth formal analysis of differences, or indications of simulation results that reveal problematic behaviour. Unfortunately, this work has to be postponed to future research.

[2] The code in the Wolfram Language is available at https://github.com/debatelab/remoma.

[3] Python Software Foundation, https://www.python.org, for a manual, see (Van Rossum and Drake, 2009).

position). An attractive feature of objection-oriented programming is inheritance, which allows for extensibility and re-usability. For example, having implemented an RE class for the default model with the quadratic function G inside the desiderata measures, a linear modal variant can inherit everything from this parent class (e.g., rules for adjustment during an equilibration process) only requiring small changes to the methods that measure the desiderata.

With a continually growing codebase we cannot rule out with certainty that the code is bug free. However, a suite of handwritten unit tests, which serves to ensure that the implementation behaves as expected, has been run over and over again.

The code of the Python implementation is available at https://github.com/debatelab/tau (classes for dialectical structures) and https://github.com/debatelab/rethon (RE classes). The code used to generate the data, raw data, and notebooks for exploration in this project are available at https://github.com/free-flux/virtuously-circular.

I intentionally keep this introduction very short even though there are countless interesting detail in the program code, e.g., improvements over the brute-force search for global optima. Such excursions would surely be interesting, but they would divert from the more philosophical aspects of this project.

9.2 Replicating Published Results

The aim of this section is to check, whether the Python implementation of the formal model yields sufficiently similar results to those which have been produced by the Mathematica implementation that accompanies the paper of Beisbart, Betz, and Brun (2021). This serves as a nice introduction to conducting ensemble studies and as an illustration of how RE simulations on a computer need to be set up. Moreover, replication is in itself a worthwhile endeavour of scientific inquiry.[4]

The dialectical structure, which serves as an example in (Beisbart, Betz, and Brun, 2021) has been introduced earlier (see Figure 7.1). They study four cases in the standard example, which are given by four different sets of initial commitments (see Table 7.1). Given that the model variant and penalties for

[4]In the present context "replication" means that we reach sufficiently similar results with different code, which implements the same formal model, and which is applied to the same dialectical structure. Variation arises from the random sampling of initial commitments and configurations of weights.

the account and the faithfulness measure are held fixed, the following need to be provided to the computer program in order to run a simulation, which I call a *simulation setup*:

- a dialectical structure consisting of a sentence pool and arguments $\tau = \langle \mathcal{S}, \mathcal{A} \rangle$
- a set of initial commitments C_0
- a configuration of weights for the desiderata in the achievement function $(\alpha_A, \alpha_S, \alpha_F)$

A simulation setup represents parts of the epistemic situation of an agent at the outset of RE inquiry. An individual simulation has two main outputs: fixed points reached by an equilibration process of mutual adjustments or global optima according to the achievement function. The current Python implementation of the formal yields the outputs that have been presented earlier in Table 7.1 for these individual RE simulation setups in the standard example.

Moreover, Beisbart, Betz, and Brun (2021) find a substantial overlap of fixed points and global optima in the present example, speaking in favour of the process being instrumental towards reaching globally optimal states, which, in turn, are among the requirements for (full) RE states. This conclusion is underwritten by two ensembles:

> The first ensemble is based on a random sample of initial conditions ($N = 500$) on the basic dialectical structure defined above. Using the same values of the weights as before, we find that in 95% of all cases, the equilibration fixed points are also global optima. Of these, 75% are full RE states. The second ensemble ($N = 500$) uses the initial commitments and dialectical structure from the four illustrative cases discussed above, but randomly varies the weights within the achievement function. Given such systematic parameter perturbation, 88% of the equilibration fixed points are also global optima; and of these, 65% are full RE states. These robustness analyses show that the process of equilibration as defined in the model is likely to lead to a global optimum or even a full RE state. (Beisbart, Betz, and Brun, 2021, 455)

The 500 simulation setups for the first ensemble consist of the dialectical structure of the standard example, 500 randomly generated initial commitments, and the standard configuration $(\alpha_A, \alpha_S, \alpha_F) = (0.35, 0.55, 0.10)$.[5] 93.0% of fixed points that resulted from the simulations with the Python are global optima, and of those 74.8% are full RE states. The second ensemble consists of 500 simulation setups for the dialectical structure of the standard example, four sets of initial commitments (cases A–D), and 125 randomly generated configurations of weights. 88.0% of fixed points are global optima and 64.7% of those are full RE states. In both cases, small differences in results are within an acceptable range of what may be expected from variation that is introduced by the random generation of initial commitments or configurations of weights. Overall, I take the results to be sufficiently similar to vindicate replication.

9.3 Finding Promising Configurations of Weights

In this section, we select promising configurations of weights and complement them to enlarge the basis of simulation setups for the upcoming larger studies.

Beisbart, Betz, and Brun (2021) use $(\alpha_A, \alpha_S, \alpha_F) = (0.35, 0.55, 0.10)$ as a the standard configuration of weights, and their simulation results as well as the replication indicate that this renders equilibration processes quite successful in reaching global optima or even full RE states. Apart from its success and noting that the individual simulations of cases A–D are robust in that they yield identical results in a larger region of parameter space (Beisbart, Betz, and Brun, 2021, 452), the configuration remains unmotivated. How does this region look like, how many other regions are there, and how extensive are they? Note that other configurations of weights are not that successful: If we repeat the simulations for the first ensemble (500 random sets of initial commitments in the standard dialectical structure) with a different weighting $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.55, 0.35)$, 97.8% of fixed points are global optima, but only 5.1% of those are full RE states!

Outputs The Python implementation of the formal model is able to produces all kinds of outputs. There are two kinds of resulting states: fixed points of equilibration processes according to the adjustment rules, and global

[5] Data and notebooks for exploration are available at https://github.com/free-flux/virtuously-circular/tree/main/chapter-9/replication.

optima according to the achievement function. We can further analyse whether fixed points are globally optimal states, or whether global optima are RE states, satisfying (CCT), or even full RE states, which satisfy (FEA).

As noted before the outputs of the implemented formal model can come apart to some extent, and in varying degrees depending on the configurations of weights. This is illustrated in Figure 9.1, which is based on another ensemble from the standard dialectical structure with 500 randomly generated sets of initial commitments for two configurations of weights

$$(\alpha_A, \alpha_S, \alpha_F) = (0.35, 0.55, 0.10) \text{ and } (\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.55, 0.35).$$

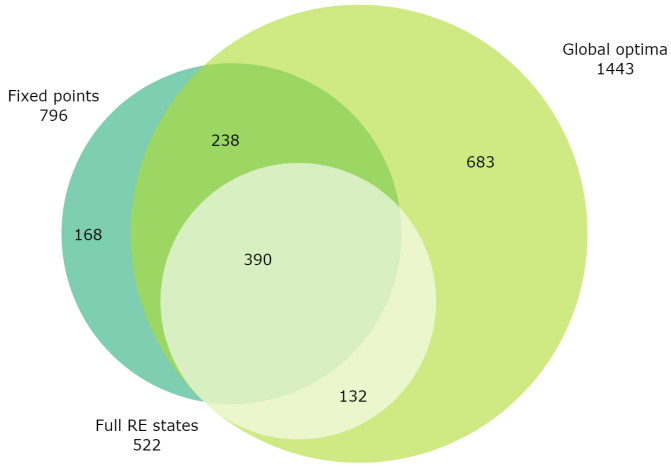
For every simulation setup, all fixed points and all global optima have been collected, as well as their status as full RE state. This allows to depict the overlap of fixed points, global optima, and full RE states in Figure 9.1.

It is important to note that the outputs are relativised to the initial commitments C_0.[6] If, for example, (C, T) is a fixed point reached from C_0 but not a global optimum with respect to $Z(\cdot, \cdot|C_0)$, it belongs to the part of the dark green circle in Figure 9.1 that does not overlap with others. However, (C, T) may well be a globally optimal fixed point with respect to a different set of initial commitments C_0', in which case the output falls in the overlap of the dark green and the light green circle.

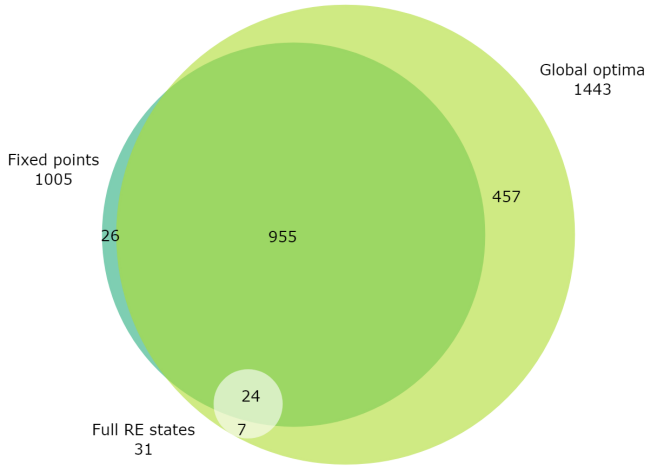
Let me add the following remarks. First, the identical numbers for global optima (1,443) for both configurations of weights is a result of the symmetry described in Proposition 9, as the configurations of weights of (A) and (B) swap the weights for account and faithfulness. Next, there is a dramatic difference between concerning the relative share of full RE states among globally optimal fixed points. For the standard weighting in Figure 9.1, (A), 62.1% of globally optimal fixed points are full RE states (390 out of 628).[7] In contrast, in (B), which differs from (A) only in the weighting but sets out from the same sets of initial commitments in the standard dialectical structure, very few simulation setups manage to yield a full RE state (31). Moreover only 24 (2.4%) of 979 globally optimal fixed points are full RE states.

[6] This is different from Proposition 3 of (Beisbart, Betz, and Brun, 2021, 467), which gets by without relativisation, and which would lead us to expect completely nested circles.

[7] This is a striking difference to the replication results. In contrast to the replication of the published results, which produced a random fixed point per simulation setup, the present ensemble collects *all* fixed points per simulation setup. Moreover, simulation setups that result in more than one fixed point exhibit a significantly lower relative share of full RE states among globally optimal fixed points. Such cases are underrepresented in the replication ensemble due to randomly choosing one fixed point.



(A) $(\alpha_A, \alpha_S, \alpha_F) = (0.35, 0.55, 0.10)$



(B) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.55, 0.35)$

FIGURE 9.1: Overlap of all fixed points (dark green), all global optima (light green) and all full RE states (white shade) from simulations in the standard dialectical structure with 500 sets of initial commitments for two different configurations of weights. The areas of circles and their intersections are in scale to the absolute numbers.

The difference between (A) and (B) in Figure 9.1 also hints at a tension. In (B), fixed points and global optima overlap to a very high degree. The relative share of global optima among fixed points is very high. In contrast, the relative share of full RE states among globally optimal fixed points is very low in (B). In turn, the configuration in (A) yields a smaller overlap of fixed points and global optima, but the relative share of full RE states among globally optimal fixed points is much higher than in (B). It seems that you cannot have all at once, i.e., fixed points that are likely to be globally optimal, and at the same time a high relative share of globally optimal fixed points that are full RE states.

We can underwrite this observation by additional weight configurations. In a next ensemble, which based on the dialectical structure of the standard example and 100 randomly chosen sets of initial commitments, we introduce more variation to the configurations of weights. The resolution of weights is $\frac{1}{25}$, resulting in $\frac{24 \cdot 23}{2} = 276$ configurations of weights without extreme values 0.0 and 1.0). This results in a total of $100 \cdot 276 = 27{,}600$ simulation setups for the standard RE model. For every simulation setup, all fixed points reached by the equilibration process were recorded, as well as whether they are global optima or full RE states.

The Relative share of full RE states among globally optimal fixed points is depicted in Figure 9.3 for every configuration of weights in a ternary plot.[8] The relative share of full RE states among globally optimal fixed points increases with the weight for account α_A. It is very low (dark) at the bottom of Figure 9.3 (mean relative share: 0.06, SD: 0.08 for $\alpha_A \leq 0.2$) and very high at the top (mean relative share: 0.95, SD: 0.09 for $\alpha_A \geq 0.6$). This may be expected as giving more weight to account makes it more likely that outputs satisfy the account-related (FEA), which is required of full RE states.

This share is notably higher for configurations of weights where $\alpha_A > \alpha_F$ (mean relative share: 0.67, SD: 0.31) in comparison to the region, where $\alpha_A \leq \alpha_F$ (mean relative share: 0.09, SD: 0.14). For a finer resolution of weights, see Figure C.1 and Figure C.2 in the appendix. Note that there is a salient "dark triangle" of extremely low relative shares at the bottom of Figure 9.3 (mean relative share: 0.02, SD: 0.01). It is defined by $\alpha_A < \alpha_F$ (angle bisector originating from the bottom left corner) and $\alpha_A < \alpha_S$ (angle

[8] For the purpose of visualisation, hexagons are used to generate a tiling of the entire parameter space by the discrete configurations of weights. They do not represent regions that achieve the same output even though this may be expected in view of Proposition 7, which states that convex regions of configurations of weights yield the same set of global optima.

bisector originating from the bottom right corner). In other words, this is a region Where faithfulness is more important than account and systematicity on their own. While these configurations of weights are favourable for producing globally optimal fixed points (see Figure 9.2), it is hindering that these globally optimal fixed points are full RE states.

Is there an explanation? $\alpha_S < \alpha_F$ is relevant in theory adjustment steps in the process of equilibration giving preference to systematic theories over theories that account well for the current commitments. In turn, $\alpha_A < \alpha_F$ pertains to commitment adjustment steps, making being faithful to the initial commitments more important than having commitments that are accounted for by the current theory. Faithfulness and systematicity can be optimised independently of each other as they pertain to different elements of the epistemic state. Thus, the resulting fixed points are likely to be globally optimal (see Figure 9.2). However, it is quite plausible that they fall short of (FEA), which involves account and which is required of full RE states. In both adjustment steps account is traded off against the other desiderata, and hence, it is not surprising that the resulting fixed points adopt theories that do not fully and exclusively account for the commitments.

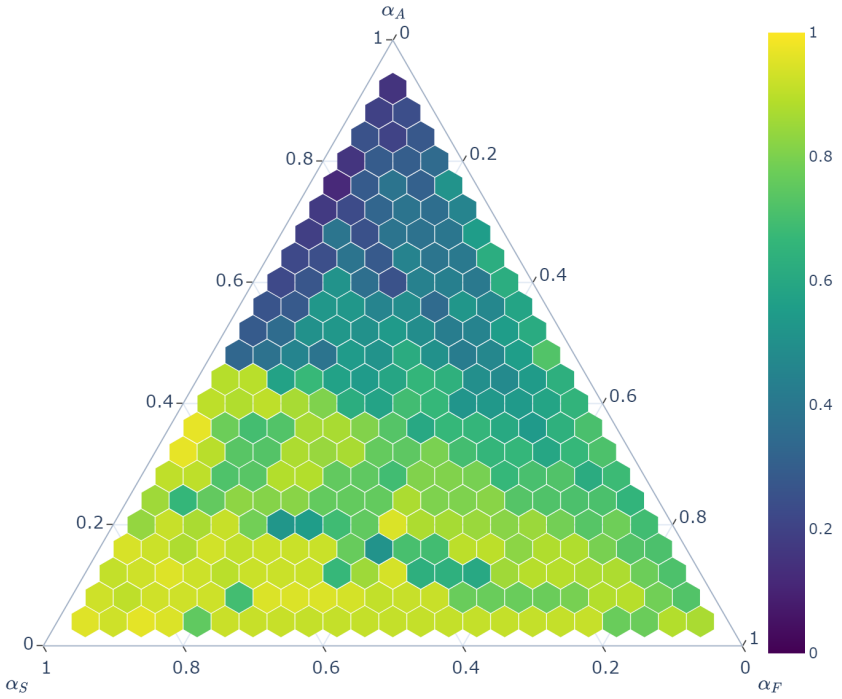


FIGURE 9.2: Relative share of global optima among fixed points reached from 100 sets of initial commitments in the dialectical structure of the standard example, grouped by configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$.

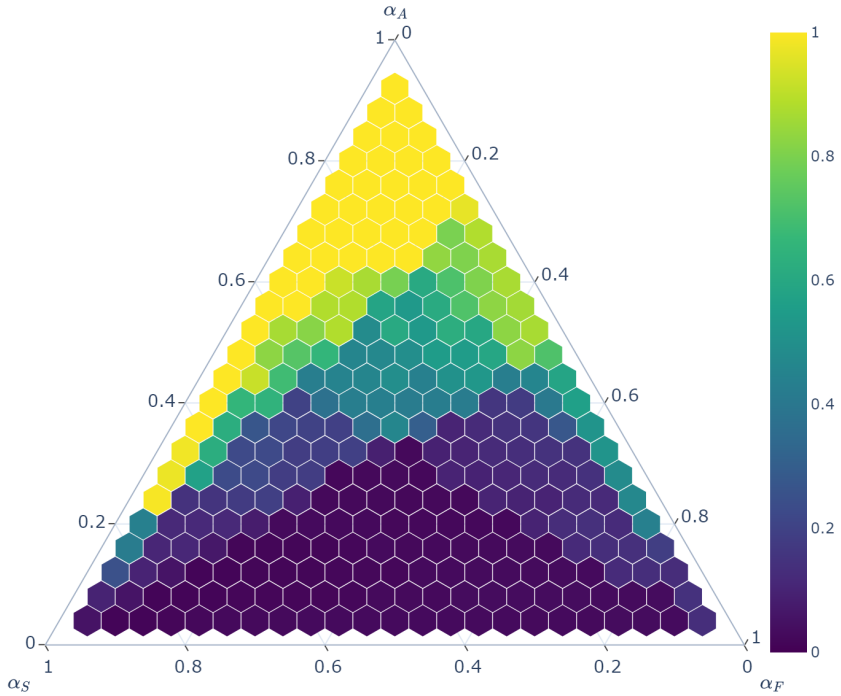


FIGURE 9.3: Relative share of globally optimal fixed points that full RE states reached from 100 sets of initial commitments in the dialectical structure of the standard example, grouped by configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$.

This raises two questions: First, which configurations of weights should we use for simulations? Configurations that yield a high overlap of fixed points and global optima or a high relative share of full RE states among globally optimal fixed points? Next, which kinds of outputs should we include in reporting the results of simulations? Fixed points, global optima, full RE states, full RE fixed points? Or should the results be reported separately for all kinds of outputs?

I will follow a two-tiered strategy. First, I will select promising configurations of weight in view of the attainment of full RE fixed points. Afterwards, I present simulation results for both global optima and fixed points separately, and irrespective of whether they are full RE states.

Why do I think that it is important to focus on full RE fixed points? A global optimum may not be a full RE state due to inconsistencies between commitments and theory, i.e., failing (CCT), or due to the theory not accounting fully and exclusively for the commitments, i.e. failing (FEA). In either case, there is room for improvement from the epistemic point of view and a consequentialist stance towards justification. The same holds for fixed points that fall short of being full RE states. In turn, full RE states that are

not fixed points may leave the proceduralist wanting as they are unattainable ideals that are not reachable by equilibration from a simulation setup (the epistemic situation). Thus, focusing on full RE fixed points for the selection of configurations aims at a synthesis between the consequentialist and proceduralist aspects of justification in RE or the selection of weights.

Afterwards, the presentation of results for global optima as well as fixed points separately allows for a series of interesting comparisons (see the remarks at the end of Section 7.1.2). First, there is the opportunity to compare results from the dynamic aspect of equilibration process, with the static aspect of states of equilibrium. Next, we also have a comparison between an unbounded agent that is able to optimise globally, and a still highly idealised, but bounded agent that optimises semi-globally. This is an intermediate step towards modelling even more rationally bounded agents that engage in RE. Moreover, we are in a position compare the results, which are justified from the proceduralist or, respectively, from consequentialist point of view about justification in RE. Finally, by not restricting our focus to only the "best" outputs of the formal model (arguably, full RE states), we can examine whether the outputs tend to perform well. In such case, it may be advisable to engage in RE to achieve some progress even if the chance of reaching a justified state of perfectly wide RE are minute. This would also stand in support of the view that RE tends to boost the epistemic standing of its inputs speaking for the justificatory power of RE. Whether this boost suffices for justification will depend on the specifics of an epistemic situation, and this cannot be investigated on the basis of randomly generated, contentless examples.

Centroids In order to extract more information about promising configuration of weights we can exploit the convexity result of Proposition 7. Full RE fixed points are global optima, and thus, they belong to convex regions of configurations of weights of the parameter space.

The fact that an individual global optimum stems from a convex set of weight configurations allows to study the "centers" of those regions as representatives. For a finite set of weight configurations

$$(\alpha_A^1, \alpha_S^1, \alpha_F^1), \ldots, (\alpha_A^k, \alpha_S^k, \alpha_F^k) \quad (k \in_{\mathbb{N}})$$

the centroid is defined as the arithmetic mean

$$\frac{1}{k} \cdot \sum_{i=1}^{k} (\alpha_A^i, \alpha_S^i, \alpha_F^i).$$

The centroid minimises the sum of squared Euclidean distances between itself and any other point in the set.

Convexity guarantees that the centroid lies inside of the set, meaning that it is a genuine representative of weight configurations that yield a specific global optimum. Figure 9.4 depicts the location of centroids in an example.

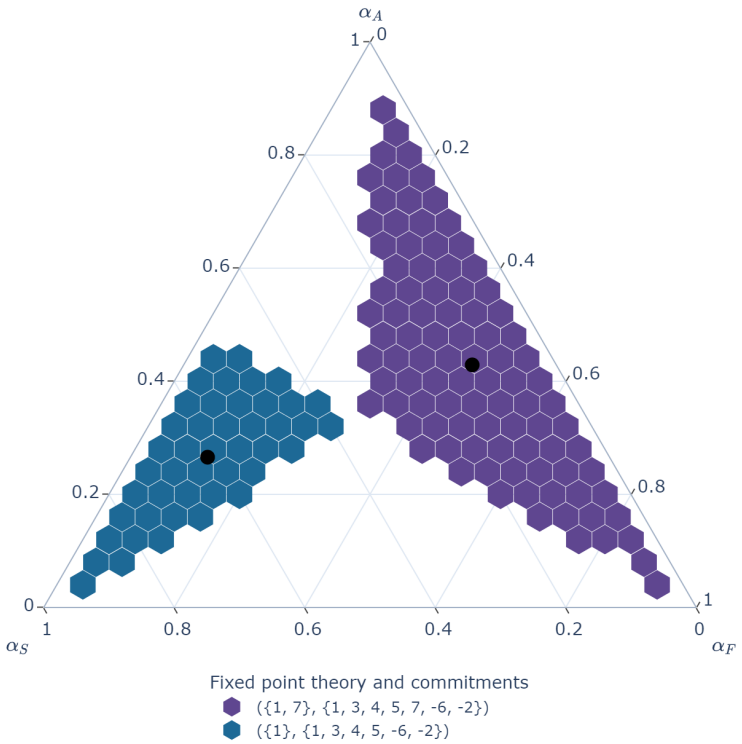


FIGURE 9.4: The black dots mark the centroids for two convex sets of configurations that yield a specific full RE fixed points in the standard example with $\{1, 3, 4, 7\}$.

The use of centroids as representatives for convex sets of weight configurations that yield a global optimum have an immediate advantage over comparing pairs of configurations in isolation. The model seems to be sensitive with respect to weight configurations that are close to, or on the border of a region. By the help of centroids as representatives, we can put this into perspective better. As "centers" of specific regions, centroids are, to some extent, more robust with respect to changes than border points. Take again Figure 9.4 as an example. The centroid of the left region (blue) is roughly at $(\alpha_A, \alpha_S, \alpha_F) = (0.26, 0.62, 0.12)$, the position of the centroid of the right region (purple) is $(0.43, 0.13, 0.44)$. The comparison reveals that there is a significant reduction in systematicity (from 0.62 to 0.13) in favour of moderate increases in account and faithfulness when we change from the blue to the purple region. Those substantial differences can explain the differences

we observe in the full RE fixed points. The full RE fixed point in the purple region is less simple (having two principles), but its commitments are more faithful to the initial commitments than the full RE fixed point in the blue region.

Figure 9.5 presents centroids of all regions of weights that yield a full RE fixed point from an equilibration process in the dialectical structure of the standard example from one of 100 randomly generated initial commitments with a weight resolution of $\frac{1}{25}$ (27,600 simulation setups). This resulted in 48,438 fixed points of equilibration processes (due to branching with random choices) of which 8,207 (17.0%) are full RE fixed points (76 unique states).

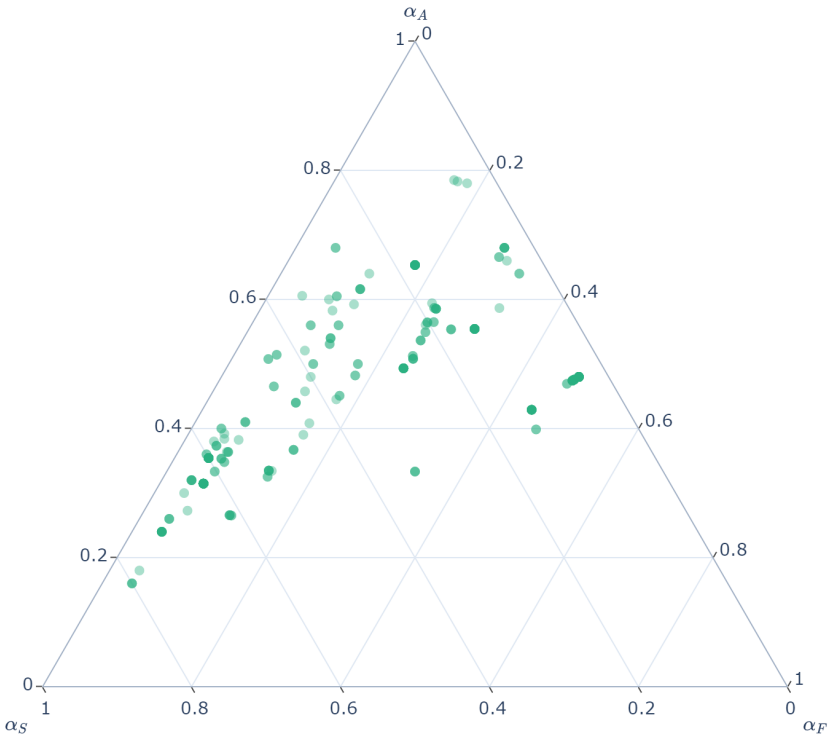


FIGURE 9.5: Centroids of regions of weight configurations that yield a full RE fixed point from an equilibration process in the standard example from one of 100 randomly generated initial commitments. Darker shades that multiple full RE fixed points and their corresponding regions resulted in the same centroid.

Most notably, there are almost no full RE fixed point centroids where $\alpha_A < \alpha_F$, i.e. the region below the angle bisector originating from the bottom left corner.[9] In addition, there are no centroids for very high values of

[9]Exceptions are (0.43, 0.13, 0.44) and (0.40, 0.14, 0.46), which slightly prefer faithfulness over account for very low weights for systematicity.

account ($\alpha_A > 0.8$). This is to be expected when taking averages. Nonetheless, it is reassuring to observe that the formal model does not achieve full RE fixed points for extreme configurations of weights.

There is also a centroid on dead centre ($\alpha_A = \alpha_S = \alpha_F = \frac{1}{3}$). It arises from "perfect" optima that cover the entire parameter space.[10] Taking the average over all configurations of weights results in this centroid.

In summary, there is a more or less distinct region of the parameter space of weighting that proves to be conducive to yielding full RE fixed points.

More Variation for Robustness Until now, only one dialectical structure was scrutinised by simulations that varied initial commitments and configurations of weights. What if completely different configurations of weights prove to be successful in different setups of dialectical structures and initial commitments? In other words, is the model robust with respect to some more variation? Overly sensitive behaviour would speak against the idea that there are configurations of theoretical virtues (and other epistemic goals), that are, in general, more plausible to yield desirable RE outputs than others.

To this purpose, we devise an ensemble of RE simulations with varying dialectical structures, initial commitments and configurations of weights:

- Sentence pool size: 7
- 10 randomly generated dialectical structures with 1–2 premises per argument, and inferential density between 0.15 and 0.5.[11]
- 100 random sets of initial commitments per structure
- A resolution of weights of $\frac{1}{25}$, yielding 276 configurations of weights without extreme values 0.0 and 1.0

This results in a total of $10 \cdot 100 \cdot 76 = 276{,}000$ different simulation setups for the standard RE model. The endpoints are all fixed points for every simulation setup, as well as their status as global optima or full RE states.

Overall, the average (across dialectical structures and configuration of weights) of the relative share of full RE states among globally optimal fixed points is 0.43 (SD: 0.38). This might be disappointing at first sight, but note that standard deviation is rather substantial. We can observe a salient "drop" when crossing the $\alpha_A = \alpha_F$ line towards the lower right half of the ternary plot in Figure 9.6, where faithfulness receives more weight than account. This

[10] For an example, see case (A) in Figure 8.11.

[11] For a definition of inferential density, see Appendix C.1.

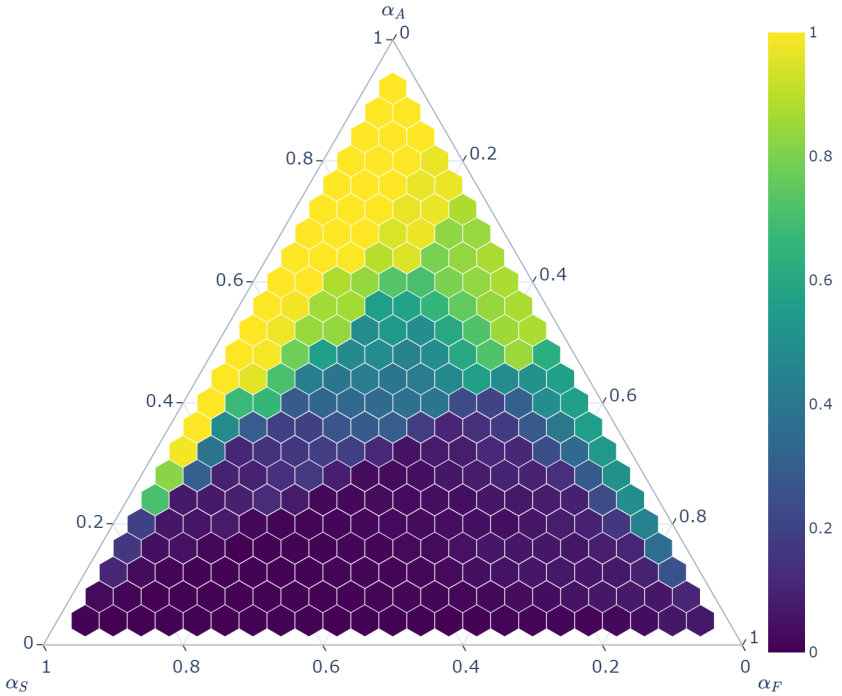


FIGURE 9.6: Relative share of globally optimal fixed points that full RE states. Avarage accros 10 randomly generated structures with each 100 sets of intial commitments.

is neatly captured taking averages over those regions. For $\alpha_A > \alpha_F$ the mean relative share of full RE states among globally optimal fixed points is 0.74 (SD: 0.27), for $\alpha_A \leq \alpha_F$ it is 0.13 (SD: 0.16). For more drastic trade-offs in favour of account, e.g., $\alpha_A > 3 \cdot \alpha_F$ (corresponding to the second salient line in the upper half) the relative share rises again to a mean relative share of 0.93 (SD: 0.13). Note that the standard weighting $(\alpha_A, \alpha_S, \alpha_F) = (0.35, 0.55, 0.1)$ is part of this last region.

Figure 9.7 depicts the centroids of all regions of weights that yield a full RE fixed point from an equilibration process the present ensemble of simulation. The 276,000 simulation setups resulted in 674,297 fixed points of equilibration processes (due to branching with random choices) of which 77,526 (11.5%) are full RE fixed points (548 unique states).

The results found in the standard example concerning the relative share of global optima among fixed points and the relative share of full RE states among globally optimal fixed points prove to be robust with respect to varying the dialectical structure in addition to initial commitments and configurations of weights. Robustness considerations with respect to variations in the sentence pool are relegated to Appendix C.3.

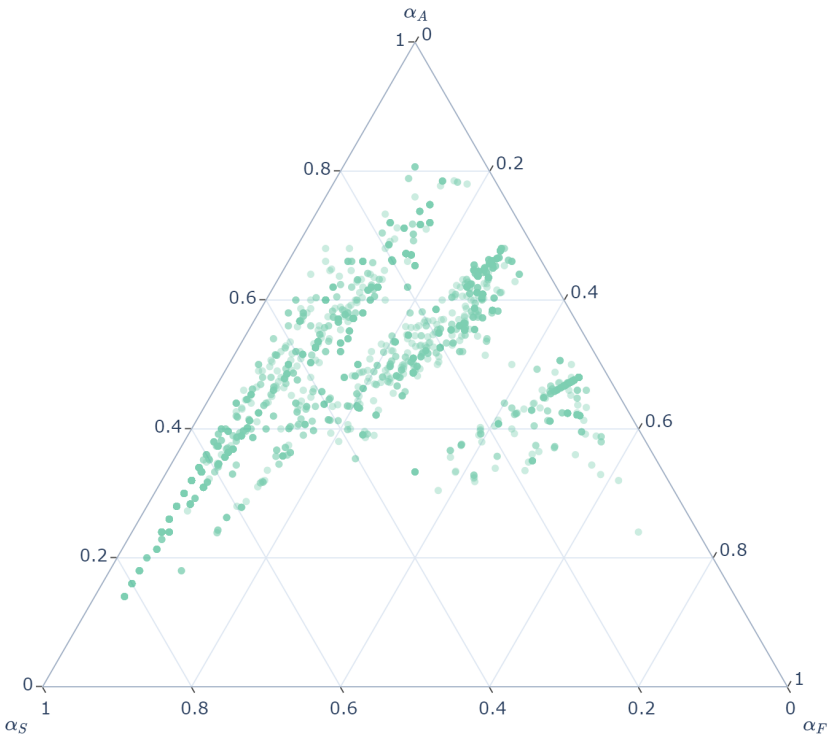


FIGURE 9.7: Centroids of regions of weight configurations that yield a full RE fixed point.Darker shades indicate that multiple full RE fixed points and their corresponding regions resulted in the same centroid.

Selecting Promising Configurations On the basis of Figure 9.8, I proceed to select configuration of weights from visual clusters of centroids that yield full RE fixed points.[12]

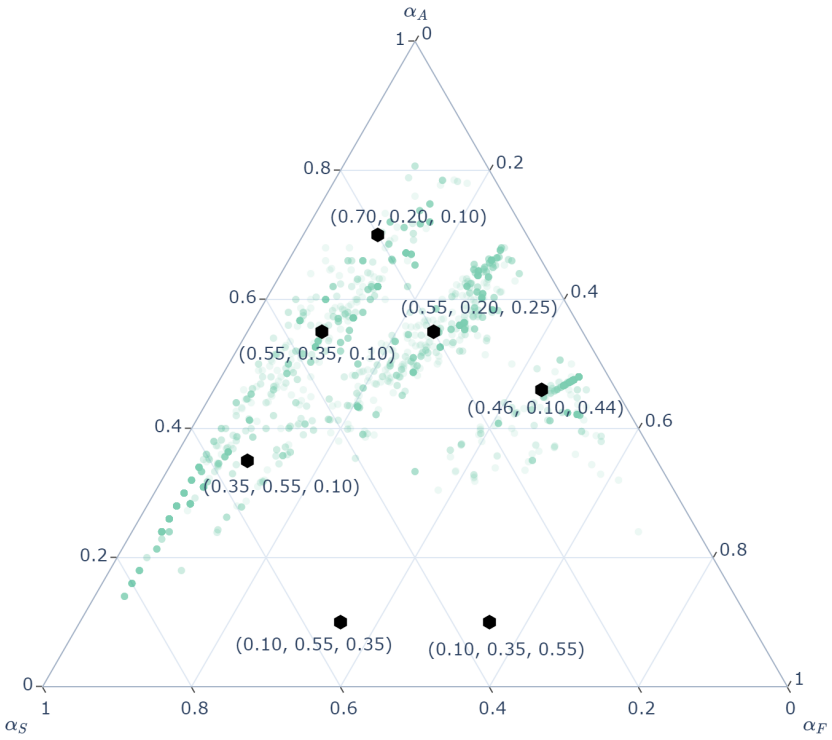


FIGURE 9.8: Selection of configuration of weights from visual clusters of full RE fixed point centroids. $(0.10, 0.55, 0.35)$ and $(0.10, 0.35, 0.55)$ complement the selection, and are taken from a region of parameter space that does not contain full RE fixed point centroids.

(0.35, 0.55, 0.10) Standard weighting used in (Beisbart, Betz, and Brun, 2021). A lot of weight is on systematicity.

(0.55, 0.35, 0.10) Swaps the weights for account and for systematicity in the standard weighting.

(0.70, 0.20, 0.10) A lot of weight is on account, and the trade-off with systematicity is more pronounced than in the previous weighting.

(0.55, 0.20, 0.25) In the centre of a pronounced cluster in Figure 9.8. Faithfulness is slightly preferred over systematicity.

[12]Clustering could also be achieved algorithmically, e.g., by DBSCAN, but this would border upon over-engineering in view of the purpose of the present project.

(0.46, 0.10, 0.44) Taken from a visual cluster of full RE fixed points from the right side of Figure 9.8 with a very low weight for systematicity. Account is granted a little more weight than faithfulness to break ties in a way that is beneficial to yield full RE fixed points (see Figure 9.3).

This selection of promising configurations of weights is complemented by two additional configurations of weights from a region of the parameter space that rarely yield full RE fixed points (see Figure 9.3) and contain no full RE fixed point centroids: (0.10, 0.55, 0.35) and (0.10, 0.35, 0.55). They serve to provide us with contrasting results in the following chapters, which address the conservativity and no-convergence objection to RE on the basis of simulation, respectively.

Appendix

C.1 Inferential Density

The number of complete and consistent extensions of a position P is denoted by σ_P, and σ_τ symbolises the number of all complete consistent positions on a dialectical structure τ. The *inferential density* of a dialectical structure is calculated as follows (Betz, 2012, 44):

$$D(\tau) = \frac{n - \log_2(\sigma_\tau)}{n},$$

where n is the size of the unnegated half of the sentence pool of τ. We have $D(\tau) \geq 0$ and the more inferential relations are imposed on τ, the less complete and consistent extensions (maximum 2^n positions for no inferential relations) are left in τ resulting in low values for σ_τ, and consequently, high values of $D(\tau)$. For the course of this project, randomly generated dialectical structures are ensured to exhibit moderate values for inferential density.

C.2 Fine-Grained Weight Resolution

The following ensemble of simulation provides a more fine-grained resolution of weights for Figure 9.2 and Figure 9.3.

- Dialectical structure of the standard example

- 40 random sets of initial commitments
- Resolution for weights: $\frac{1}{50}$ yielding 1,176 configurations of weights

This results in 47,040 simulation setups, for which all fixed points as well as their status as global optima and as full RE states have been recorded. The results are depicted in Figure C.1 and in Figure C.2 with more detail. There are no notable differences to the original, coarse-grained results in 9.2 and in Figure 9.3. As 40 new random sets of initial commitments have been used for the simulations, the findings reported for the coarse-grained ensemble above are corroborated.

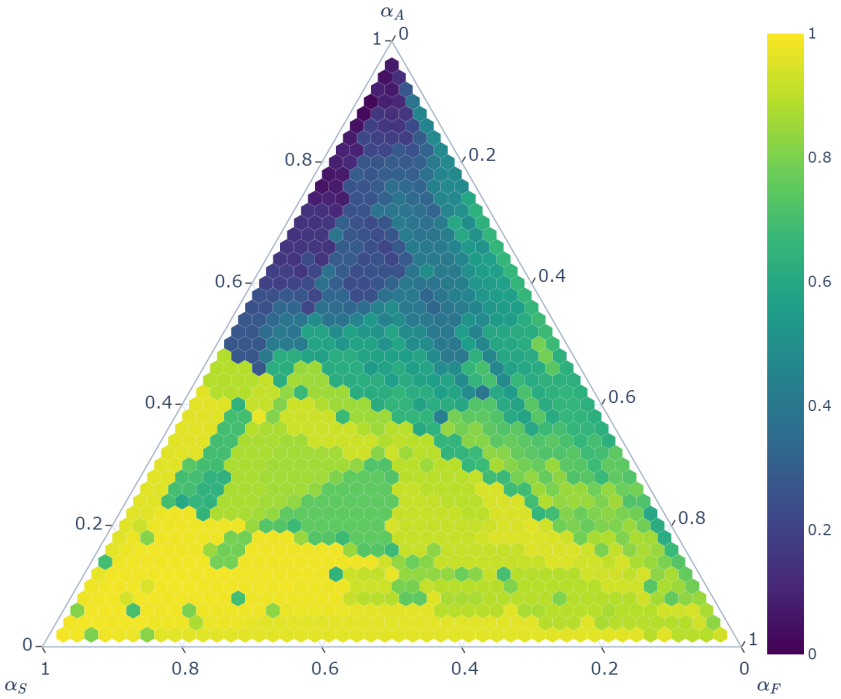


FIGURE C.1: Relative share of global optima among fixed points reached from 40 sets of initial commitments in the dialectical structure of the standard example.

C.3 Varying Sentence Pool Sizes

The ensemble to check robustness with respect to varying sentence pool sizes covers 6 to 9 unnegated sentences. For each pool size, 10 randomly generated dialectical structures (inferential density between 0.20 and 0.50), and for each structure, 4 random sets of initial commitments. The weight resolution is $\frac{1}{50}$ yielding 1176 different configurations of weights without extreme values 0.0

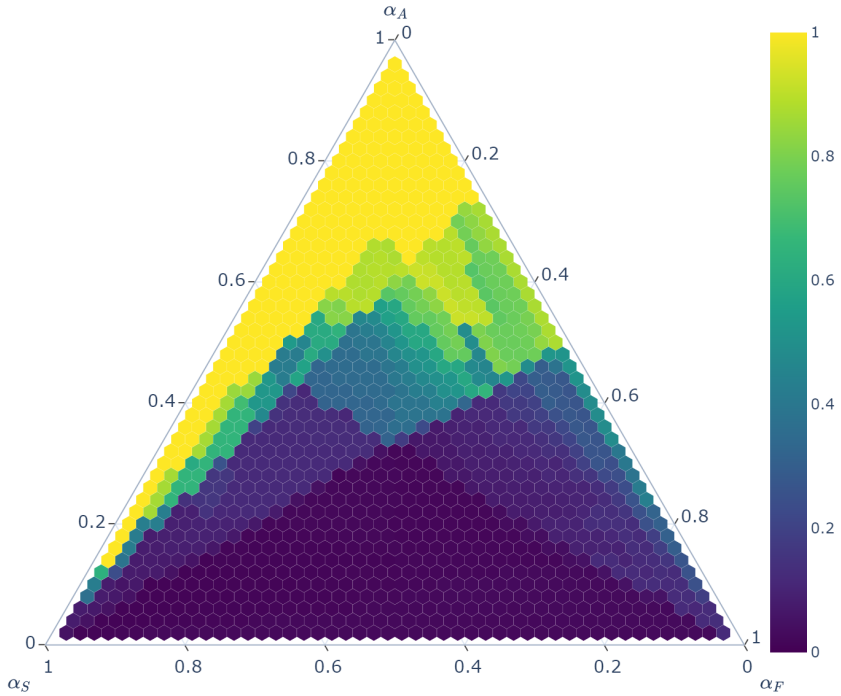


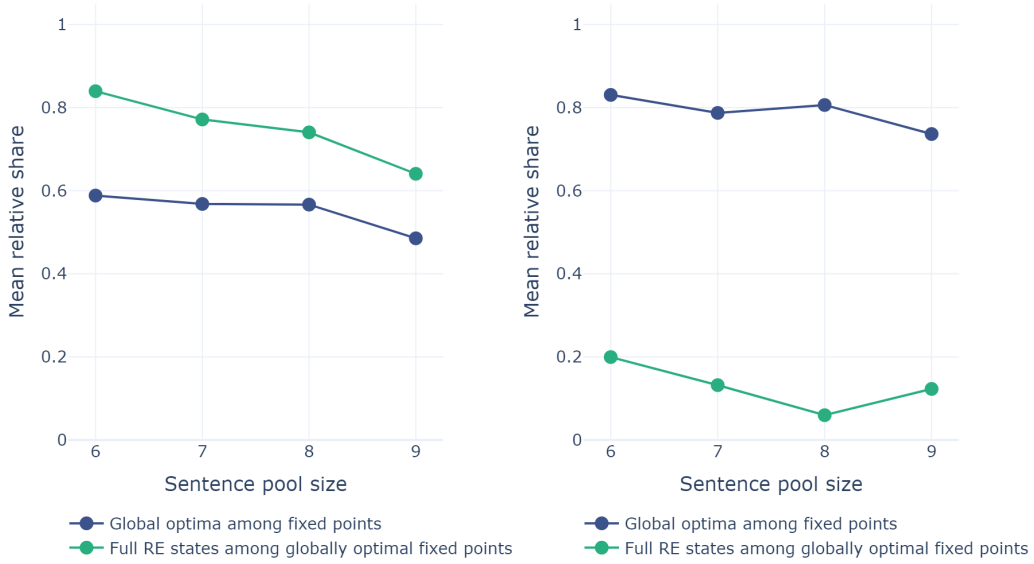
FIGURE C.2: Relative share of globally optimal fixed points that full RE states reached from 40 sets of initial commitments in the dialectical structure of the standard example.

and 1.0. This amounts to 188160 simulation setups for the implementation of the standard model. The endpoint of each is a single fixed point of the equilibration process (without tracking other branches due to random choices), as well as its status as global optimum and full RE state.

Figure C.3 depicts the relative share of global optima among fixed points and full RE states among globally optimal fixed points, grouped by sentence pool size. The stark difference between regions (A) and (B) is in line with observations that we made earlier in ternary plots, e.g., for Figure 9.2 and Figure 9.3.

The relative shares tend to slightly decrease with larger sentence pool sizes. It is an open question at which point we should deem the relative share of full RE states among globally optimal fixed points unacceptably low for the implementation of the formal model (if it even is computationally feasible). From an informal point of view it seems plausible, that to achieve a state of RE gets more difficult the more we take into consideration, and ultimately rendering perfectly "wide" RE an unattainable ideal.

For now, another robustness result is more important. The same regions



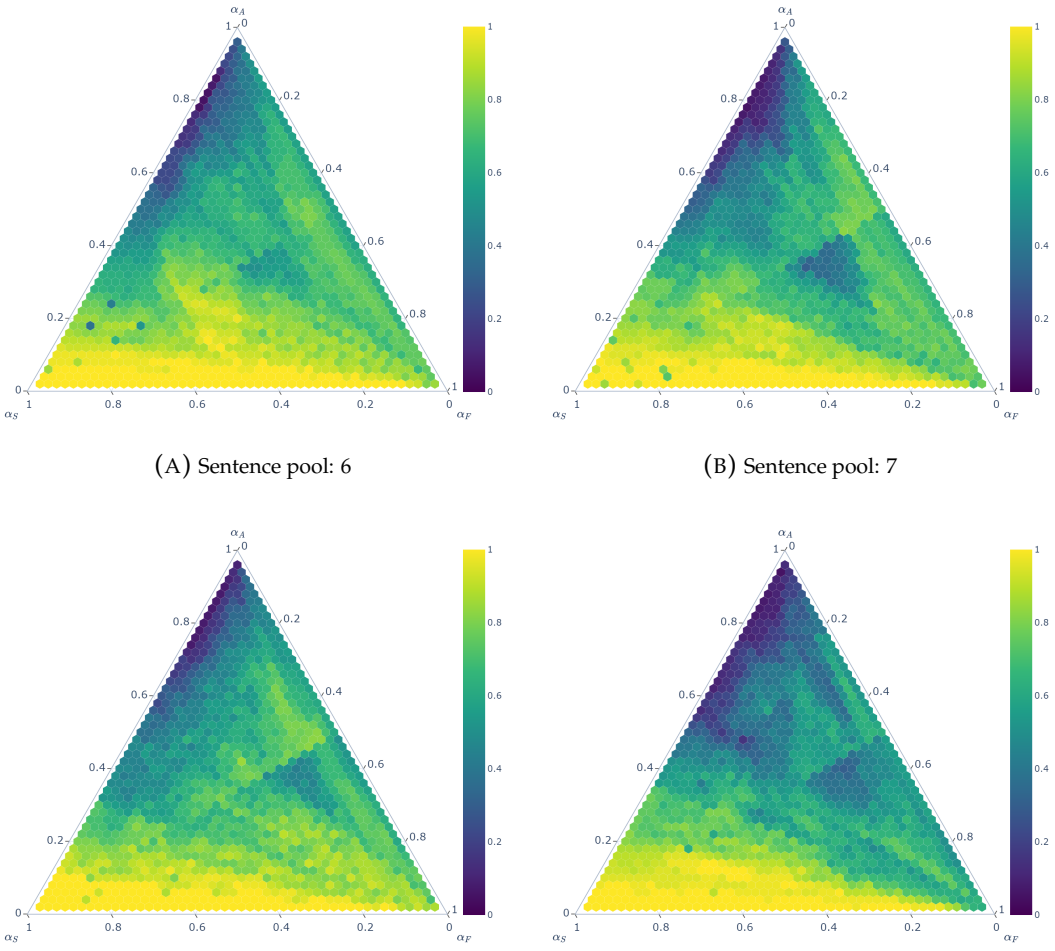
(A) $\alpha_A > \alpha_F$ (B) $\alpha_A \leq \alpha_F$

FIGURE C.3: Relative shares of specific outputs (average across two specific regions of the parameter space) grouped by sentence pool size.

of configurations of weights prove to be conducive to yield specific RE outputs independent of the sentence pool. This is illustrated in Figure C.4 (global optima among fixed points), Figure C.5 (full RE states among globally optimal fixed points) ,and Figure C.6 (full RE fixed point centroids).

Figure C.6 depicts the centroids of regions of weights that yield a full RE fixed point for different sentence pool sizes. There are more individual centroids for higher sentence pool sizes[1] accompanied by increased spread towards the top of plots (more weight on account). Aside from that, there are also robust findings. There is some concentration of full RE fixed point centroids on the bottom left side of plots, where faithfulness is very low, and systematicity receives more weight. Note that this is roughly the region from where the standard weighting $(\alpha_A.\alpha_S, \alpha_F) = (0.35, 0.55, 0.1)$ stems. Furthermore, the region where account receives less weight than faithfulness contains almost no centroids irrespective of the sentence pool size.

[1]For sentence pool sizes 6, 7, 8, and 9 there are 124 137, 141, and 152 centroids, respectively.



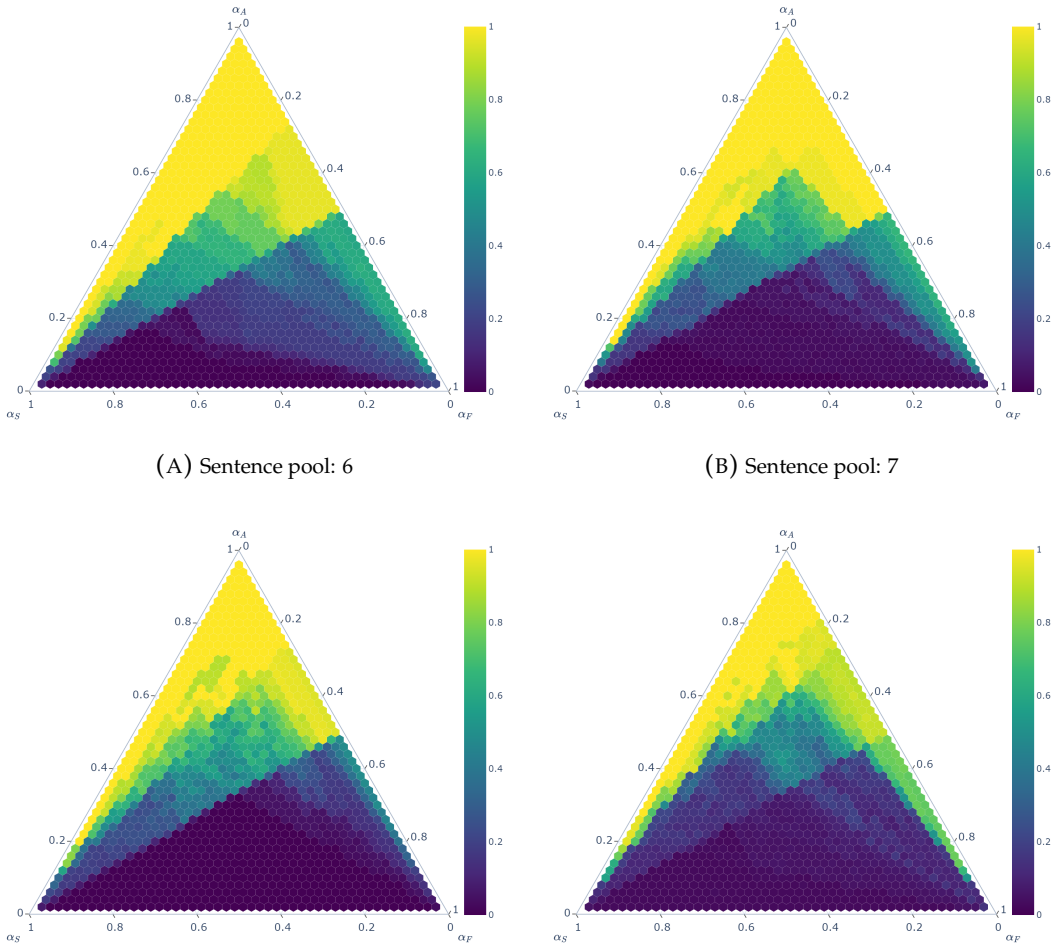
(A) Sentence pool: 6

(B) Sentence pool: 7

(C) Sentence pool: 8

(D) Sentence pool: 9

FIGURE C.4: Relative share of global optima among fixed points for different sentence pool sizes. The results are in line with findings in Figure 9.2 and in Figure C.1



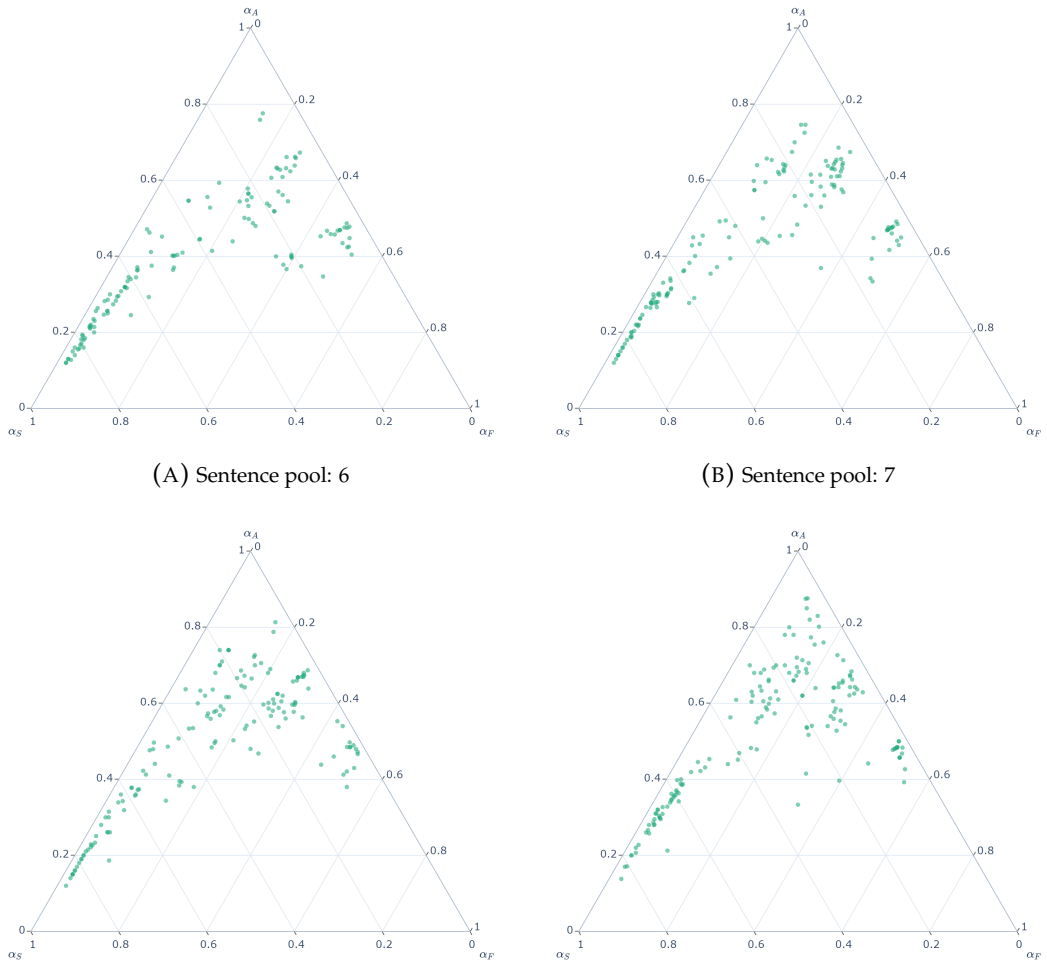
(A) Sentence pool: 6

(B) Sentence pool: 7

(C) Sentence pool: 8

(D) Sentence pool: 9

FIGURE C.5: Relative share of full RE states among globally optimal fixed points for different sentence pool sizes. The results are in line with findings in Figure 9.3 and in C.2



(A) Sentence pool: 6

(B) Sentence pool: 7

(C) Sentence pool: 8

(D) Sentence pool: 9

FIGURE C.6: Centroids of regions of configurations of weights that yield a full RE fixed point for different sentence pool sizes. They serve as a robustness analysis of findings for Figure 9.5 and Figure 9.7.

Chapter 10

Is Reflective Equilibrium Too Conservative?

10.1 Introduction

Recall the suspicion of conservativity directed against RE from Chapter 3:

(Conservativity) RE does not provide enough incentive for a substantial revision of initial commitments.

This is problematic for justificatory power of RE, if the agents set out from epistemically defective starting points. If the outputs preserve the deficiencies, they cannot be justified. Such outcomes would stand in support of (Weakness), the claim that being in a state of RE is not sufficient for justification. Against the view that RE merely streamlines the initial commitments with minor adjustments, proponents of RE suppose that systematisation is the "key driver" behind an equilibration process of more thorough revisions. Thus, the involvement of theoretical virtues in systematisation may be relevant to ward off the threat of conservativity.

The aim of this chapter is to address the question, whether RE is too conservative on the basis of data generated by the computer implementation of the formal model of Beisbart, Betz, and Brun (2021). As it stands, (Conservativity) targets the informal accounts of RE, and hence, some preparatory work is required to operationalise (Conservativity) for examination in the formal model. The formal model allows to operationalise three important aspects that are relevant to the discussion of conservativity in RE. We can present them as questions: Does RE lead to substantial change? Does RE dispose of garbage? Does RE make views more systematic? (Conservativity) would let us expect that the answer to these questions is a resounding no.

The chapter is organised as follows: In Section 10.2, I introduce a streamlining procedure to have a conservative baseline for comparison with behaviour of the formal model, and I provide general information about the generation of data. The following sections (10.3, 10.4, and 10.5) treat individual aspects of conservativity under the rubric of background-method-result-discussion. I keep the background at a bare minimum due to the more detailed, informal treatment in Chapter 3. Section 10.6 concludes the chapter with a summary of general study results and their repercussions for RE and theoretical virtues. I relegate robustness consideration in view of varying configuration of weights and sentence pool sizes to Appendix D.1.

10.2 Preparations

10.2.1 A Streamlining Baseline

When we ask whether RE is too conservative, we need to specify what "too conservative" means. Two approaches are apparent. On the one hand, we could set a threshold for what would count as sufficiently non-conservative behaviour. For example, we might reject (Conservativity) with respect to inconsistency preservation if more than half of inconsistent initial commitments resulted in consistent output commitments.

An immediate issue of this approach is the question whether we can give any motivating reasons for choosing a specific threshold over another to mark off conservative behaviour. Otherwise, the threshold may appear arbitrary and this impression may be worsened if the general tendencies in the results are not clear-cut cases. Why settle for an unambitious threshold at 0.5 and not, say, 0.75 or even 0.95?

On the other hand, we can study whether RE performs better than a baseline. During the informal presentation of objections to RE (Section 3.2), I described a streamlining procedure informally. It is conservative by design, and its operationalisation in the formal framework will serve as baseline for comparisons with the outputs of the formal model.

In my view, the baseline approach is more suitable than thresholds. First, we can escape the worrisome arbitrariness of having to specify thresholds. Next, the streamlining baseline takes up the simplistic depiction of RE, which is passed around in the literature, that a state of coherence can be reached with minimal effort. Finally, the baseline is almost devoid of theoretical virtues, and does not allow for trade-offs between them. Consequently, the

baseline procedures are also a helpful tool to study the influence of theoretical virtues implemented in the formal model on conservativity.

I need the following definitions to operationalise the streamlining procedure in the framework of the formal model of RE. Let us assume that a dialectically structure τ is given. An *axiomatic base* of a dialectically consistent position P from a source of positions $\mathfrak{S}$ is a position $Q \in \mathfrak{S}$, such that Q entails P and there is no strict subset of Q that entails P. A position Q dialectically *entails* another position P if and only if every consistent and complete position that extends Q also extends P. Equivalently, Q entails P if and only if P is a subset of the dialectical closure $\overline{Q}$. An axiomatic base Q of P from source $\mathfrak{S}$ is called *minimal* if there is no other axiomatic base of P from source $\mathfrak{S}$ that contains strictly less elements than Q. The *remainders* of a position P are maximal sub-positions of P that are dialectically consistent. For a dialectically consistent position the set of remainders contains only the position itself.

Here are the instructions for an agent to apply the operationalised version of the streamlined procedure (for schematics, see Section 3.2): Start with a set of initial commitments C_0. If they are inconsistent, choose a remainder as your current commitments C. Otherwise, keep $C = C_0$. Find the minimal axiomatic base T for your current commitments C from the source of all sub-positions of C. Adjust your commitments to the dialectical closure of the axiomatic base $C' = \overline{T}$. Stop with (C', T).

Streamlining is conservative by design in the following sense: As only axiomatic bases from a severely restricted source can serve as theory candidates, changes in the commitments are kept at a minimum while guaranteeing dialectical consistency. Streamlining blocks the selection of more systematic or better fitting theories that would have higher revisionary potential.

As there may be multiple remainders for a dialectically inconsistent position, and multiple minimal axiomatic bases for a consistent one, the streamlining procedure can yield multiple outputs. Analogous to branching equilibration processes, we resolve such cases by a random choice, and keep record of every path that the streamlining procedure can take.

For an example, take $C_0 = \{2, 3, 4, 5\}$ in the standard example of (Beisbart, Betz, and Brun, 2021) from Section 7.1.2. The initial commitments C_0 are dialectically inconsistent (2 entails -4), and its remainders are $\{3, 4, 5\}$ and $\{2, 3, 5\}$. For the former remainder, the minimal axiomatic base (given the source of all sub-positions of $\{3, 4, 5\}$) is again $\{3, 4, 5\} = T$, and its dialectical closure is $\{-2, 3, 4, 5\} = C$. The streamlining procedure terminates

with (C, T). For the latter case, the axiomatic base of $\{2, 3, 5\}$ is $\{2, 3\} = T'$. The dialectical closure of T' is $\{-1, 2, 3, -4, 5, 6\} = C'$. Consequently, the streamlining procedure results in (C', T').

Note that this result differs significantly from the formal model, where $T = \{1\}$ is part of a unique global optimum (and fixed point) reached from $C_0 = \{2, 3, 4, 5\}$ (c.f. Beisbart, Betz, and Brun, 2021, 452). In contrast, the streamlining procedure cannot select $T = \{1\}$, as 1 is not in any subset of C_0.

With respect to theoretical virtues, the streamlined outputs are interesting, as they incorporate consistency and full and exclusive account as nearly sole virtues in the procedure. Note that both of them are categorical virtues. The selection of maximally consistent subsets and minimal axiomatic bases involve some kind of optimality, but we can think of them as basic procedural requirements of rational belief change rather than theoretical virtues.[1] Consequently, the streamlined outputs also serve as a baseline to study the influence of additional virtues implemented in the default model, especially the interplay of gradual virtues of simplicity, scope and account.

10.2.2 General Information About Simulations

The data for this study stems from the following setup: The sentence pool size was fixed at 7 unnegated elements and 100 dialectical structures were randomly generated. [2] In every dialectical structure, 25 randomly chosen, minimally consistent positions served as initial commitments. The selection of configuration of weights comprises the seven elements from Section 9.3. This results in $100 \cdot 25 \cdot 7 = 17{,}500$ simulation setups. For each simulation setup we keep track of all reached fixed points, global optima, and outputs from the streamlining procedure.[3]

I focus on the relation of initial and output commitments to operationalise the aspects of conservativity. In view of my preoccupation with theoretical virtues in RE, this might be a surprising move. After all, theories are supposed to be the bearers of theoretical virtues. The reason for this move is

[1] The present definition of remainder is inspired by Belief Revision Theory (e.g., Hansson, 1999, 12). The basic operations of rational belief change, contraction and revision, consist of selections among maximal subsets of a set of sentences that do not imply a specific belief.

[2] The method that randomly generates dialectically structures ensures that the resulting structures are minimally orderly. The arguments are jointly satisfiable, they are not question-begging by repeating the conclusion among premises, they avoid flat contradictions, and they do not use the same premises for different conclusions.

[3] Data sets, as well as interactive notebooks for exploration are available at https://github.com/free-flux/virtuously-circular/chapter-10.

rather mundane: The informal discussion of RE does not consider initial theories, and the formal model of RE defaults to the empty position $T_0 = \emptyset$. So, if there are no interesting initial theories, there is nothing that could be preserved in the output.[4] In addition, an inspection of the ensemble generated for the current study reveals that the theory is a subset of the output commitments in most cases. The mean relative share of global optima (fixed points) commitments containing the theory as a subset is 0.92 (0.92). Thus, whatever virtuous features are present in an output theory, they are almost always incorporated into the output commitments, too.

A simulation is individuated by the dialectical structure, the initial commitments and a configuration of weights for the achievement function. I group simulations by their configuration of weights averaging over dialectical structures and initial commitments to bring the influence of trade-offs between theoretical virtues to the fore. I restrict the presentation of study results to two paradigmatic configurations that yield very different outcomes: (A) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$, which gives a lot of weight to account and some to systematicity. In contrast, (B) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$ gives a lot of weight to faithfulness. I relegate additional configuration of weights to the appendix D.1.1.

10.3 Does RE Lead to Substantial Change?

Background Conservativity is coupled with the idea of minimal change summarised as a slogan: "If you have to change something, change as little as possible". A quick gloss over the literature in in Chapter 3 revealed that this is a popular line of thought pursued in many fields from epistemology and philosophy of science to Belief Revision Theory.

Against this backdrop, detractors of RE raise the suspicion that agents give undue weight to their initial commitments, and that RE does not provide enough incentive to revise them substantially. Allegedly, minor adjustments suffice to establish coherence, which is often characterised in terms of consistency and fit between commitments and theory. At this point, the

[4] The dialectical structures form the background of inquiry in the formal model, and they are fixed during equilibration or global optimisation. Informally, background theories are not immune to revision, and hence, may also be subject to change. However, this would put the background theory in the foreground of another RE inquiry. For a more detailed treatment of foreground and background, see Baumberger and Brun (2021).

trigger for further revision, incoherence, vanishes. Against this view, proponents of elaborate accounts of RE take systematisation to bear the potential for more thorough revisions.

On the most basic level, we can examine differences on the level of sentences between initial and output commitments. From the viewpoint of (Conservativity), only minor changes in terms of elements from initial to endpoint commitments are to be expected.

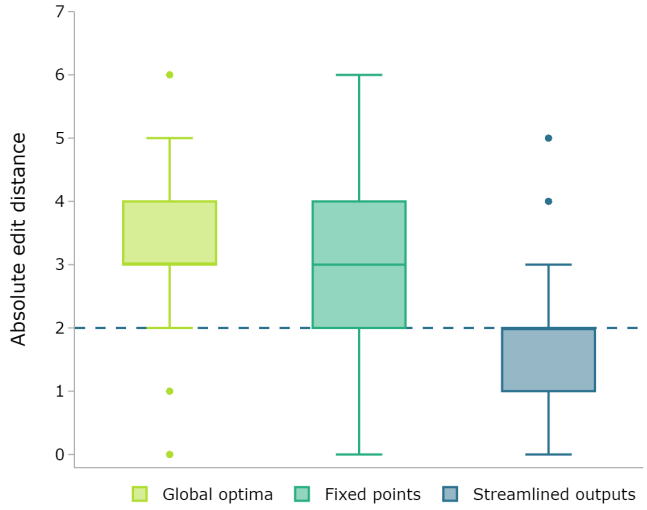
Method The generalised Hamming distance between two positions (sets of sentences) conveys a simple idea of how much changed from initial to output commitments on the level of sentences. As the sentence pool of a dialectical structure is fixed, we can sum over all sentences and penalise sentences for which the positions differ. Let $\mathcal{S}$ be a sentence pool with n unnegated sentences s_i ($i = 1, \ldots, n$), and let P and Q be positions built up from sentences from $\mathcal{S}$. Recall that the *Hamming distance* between P and Q from Section 7.1.2:

$$d_{d_0,d_1,d_2,d_3}(P, Q, \{s_i, \neg s_i\}) = \begin{cases} d_3 & \text{if } \{s_i, \neg s_i\} \subset (P \cup Q) \text{ (contradiction)} \\ d_2 & \text{if } \{s_i, \neg s_i\} \cap (P) \neq \emptyset \\ & \text{and } \{s_i, \neg s_i\} \cap (Q) = \emptyset \text{ (contraction)} \\ d_1 & \text{if } \{s_i, \neg s_i\} \cap (P) = \emptyset \\ & \text{and } \{s_i, \neg s_i\} \cap (Q) \neq \emptyset \text{ (expansion)} \\ d_0 & \text{otherwise (agreement)} \end{cases}$$

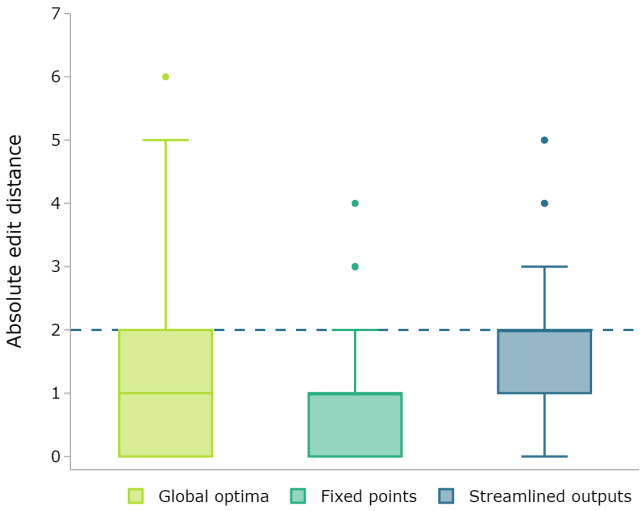
If the penalties are set to $d_3 = d_2 = d_1 = 1$, and $d_0 = 0$, we effectively count the differences between two positions, which corresponds to the number of individual edits to transform one position into the other. This edit distance between initial and output commitments is the study endpoint. Here, I refrain from normalisation by the size of the sentence pool to convey an intuitive idea of how many sentences have been altered. Note that the measure of faithfulness $F(C|C_0)$ from the default model also gives us an idea of "farness" from the initial commitments, but it does not capture the distance between initial and endpoint commitments because expansions are not penalised in the measure of faithfulness ($d_3 = d_2 = 1, d_1 = d_0 = 0$).

Results Figure 10.1 displays the distributions of absolute edit Hamming distances between initial and output commitments. As this figure condenses a lot of information, I go through it in more detail to offer some clues about how to interpret subsequent figures that are similar. The two subplots, (A)

and (B), correspond to two configurations of weights that cause the model to behave quite differently.



(A) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$



(B) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$

FIGURE 10.1: Absolute edit distance between initial and output commitments. The dashed line corresponds to the median edit distance of the streamlining baseline.

Figure 10.1 depicts the absolute edit distance (numbers of sentences that changed) between initial and output commitments for model outputs (global optima: light green, fixed points: turquoise) as well as the streamlining baseline (blue) on the vertical axis. The higher, the more distance between initial and output commitments.

The figure is a so-called box plot, which is a convenient tool to condense and display distributions and important numerical features of data. The central box contains 50% of values from an ordered data set and it includes the middle value, called *median* that is represented as a solid line inside of the box. The box is restricted by the first quartile (25% percentile) from below and the third quartile (75% percentile) from above. Thus 25% of values from the data set lie below the first quartile and another 25% above the third quartile. The distance between the first and the third quartile is called *inter quartile range* (IQR). The whiskers attached to the box have a maximal length of $1.5 \times IQR$ (or are restricted to the most extreme actual values covered by them). Every value outside of the box and the whiskers is treated as an *outlier* represented by a dot.

The median of the streamlining baseline is stretched out over the entire plot as a blue, dashed line for reference. If the median distance between the initial commitments and the outputs of the formal model is lower than the median of the baseline, the RE model performs on average worse than a procedure that is conservative by design. I take this to signify overly conservative behaviour of the formal model.

In Figure 10.1 we can observe that the streamlining baseline procedure manages to change some sentences between initial and output commitments (median: 2, IQR: 1–2). The streamlining baseline does not involve a configuration of weights, and hence the results are identical for (A) and (B).

For configuration in (A), the model yields distances for global optima (median: 3, IQR: 3–4) and fixed points (median: 3, IQR: 2–4) that exceed the streamlining baseline, on average. Thus, for the configuration in (A), the formal model of RE changes more sentences between initial and output commitments than the streamlining baseline. In contrast, for the configuration in (B), the distances of global optima (median: 1, IQR: 0–2) and fixed points (median: 1, IQR: 0–1) do fall below the streamlining baseline, on average.

Discussion The results indicate the configuration of weights is relevant to the model's behaviour with respect to conservativity operationalised as little change on the level of sentences. There are configuration of weights that lead the formal model to perform better or worse than the conservative baseline of streamlining. In contrast to the streamlining baseline, including gradual desiderata for account and systematicity that allow for trade-offs, lead to more change on the level of sentences when properly weighted.

It is not surprising that giving a lot of weight to faithfulness (B) renders the model to behave overly conservative. Faithfulness depends on a weighted Hamming distance between the initial and the current commitments of an epistemic state. If faithfulness receives a lot of weight, low distances are incentivised, which leads to few changes on the level of sentences. This is reflected when we measure the edit Hamming distance between initial and output commitments.

Taking the median as well as its range into account, global optima perform slightly better than fixed points for both configurations. Plausibly, this is due to the fact that global optimisation can reach states with more vigorously revised commitments by selecting commitments and theories simultaneously. Such states may not be available in the alternating, semi-global optimisation steps during an equilibration process.

10.4 Does RE Dispose of Garbage?

Background Even in face of substantial change on the basic level of sentences, RE may still be conservative with respect to features on a higher level. Namely, on the level of positions, i.e., sets of sentences, the issue of (Conservativity) becomes more pressing in face of epistemically deficient inputs, i.e., due to bias or prejudice ("garbage in"). If such initial deficiencies are preserved through a conservative process or by a weak characterisation of RE states, the outputs are likely to be epistemically deficient as well ("garbage out").

There is no straightforward path to equip the default model with a natural measure for bias, prejudice or other forms of epistemically deficient inputs.[5] Instead, it is more appropriate to look at interesting features of positions that can be derived from a dialectical structure.

Method Dialectically inconsistent initial commitments are a prime example of epistemically deficient inputs. Does RE preserve inconsistencies from initial commitments? (Conservativity) would lead us to expect an affirmative answer.

[5] Of course, we could equip positions with a real number representing its bias or absurdness. However, such a move would dramatically increase the number of free parameters in the model, and give rise to a series of tricky questions: Are these values subjective credences? Do they obey the laws of probability? How do we determine or interpret the value of a sentence? I do not claim that such questions cannot be answered in a philosophically insightful manner, but they go far beyond the scope of the project at hand.

	output consistent	**output inconsistent**
input consistent	consistency preserving (CP)	consistency eliminating (CE)
input inconsistent	inconsistency eliminating (IE)	inconsistency preserving (IP)

TABLE 10.1: Four cases arise from the combination of consistent and inconsistent input and output commitments.

Initial commitments are dialectically consistent or inconsistent, and the same applies to output commitments. Consequently, there are four distinct cases that depend on the consistency status of initial and output commitments, which are depicted and labelled in Table 10.1.

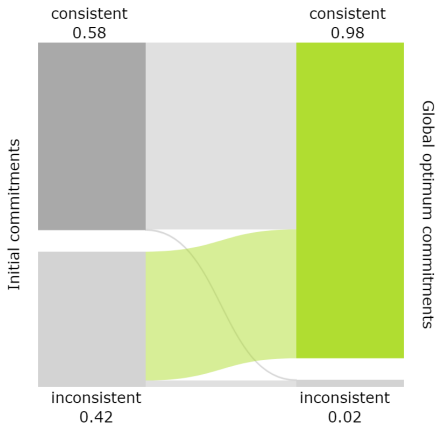
The most relevant cases for conservativity are IE and IP, where initial and output commitments are dialectically inconsistent. (Conservativity) would lead us to expect that inconsistency preserving cases (IP) occur often relative to all cases starting from inconsistent initial commitments (IP + IE). Conversely, inconsistency eliminating cases (IE) would be scarce, indicating lack of revisionary power of RE with respect to consistency.

The other cases are interesting in their own right as they serve to validate the formal model beyond the existing results in (Beisbart, Betz, and Brun, 2021). Consistency preserving cases (CP) are a benign form of conservativity as consistency is an epistemically desirable feature of input commitments. Finally, consistency eliminating cases (CE) are a troubling feature for an RE model, as even detractors of RE do not envisage such cases to occur. In summary, a positive evaluation of the default model would require a high share of (IE) and (CP) cases contrasted with low shares of (IP) cases and no occurrences of (CE) cases.

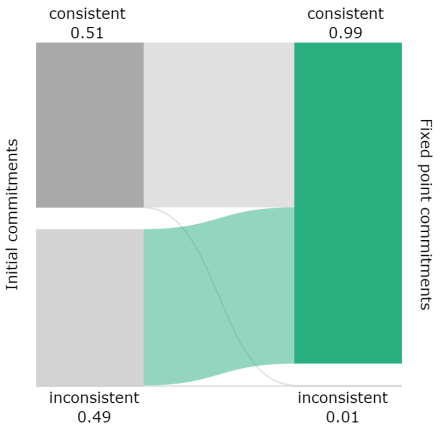
Unfortunately, the streamlining procedure cannot serve as a helpful baseline for this part of the study. It guarantees consistent output commitments, and thus, does not produce inconsistency preserving cases. Hence, the study endpoint is the relative share of consistency cases for global optima and fixed points.

Results Every subplot in Figure 10.2 depicts the relative shares of consistent and inconsistent initial commitments on the left, as well as for output commitments on the right.

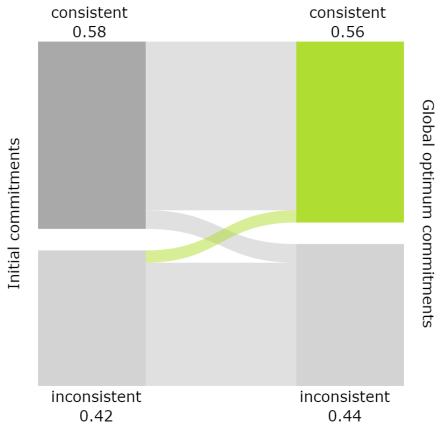
Overall, we can observe that the randomly selected initial commitments



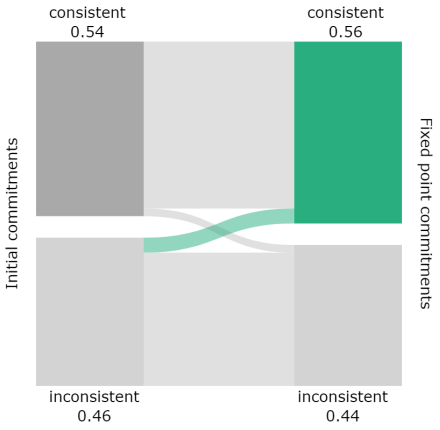
(A) Global optima
$(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$



(B) Fixed points
$(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$



(C) Global optima
$(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$



(D) Fixed points
$(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$

FIGURE 10.2: Relative shares of consistent and inconsistent initial and output commitments. Bands between the bars relate inputs to outputs according to the consistency cases. The coloured band corresponds to inconsistency eliminating case (IE), which is indicative of the formal model's revisionary power.

are consistent in roughly half of all cases (dark grey).[6] For the first configuration in (A) and (B), the relative share of consistent commitments among all outputs is substantially boosted (global optima: 0.98, fixed points: 0.99). For the other configuration, the relative share of consistent commitments among outputs remains roughly the same in (C) and (D).

The connecting bands convey an impression of the consistency cases. Inconsistency eliminating cases (IE) are coloured. There is a stark difference between the configurations. In (A) and (B), the band leading from inconsistent initial commitments to consistent output commitments is very wide. This corresponds to high relative shares of inconsistent initial commitments that result in consistent output commitments. In contrast, the coloured IE band is rather thin in (C) and (D), corresponding to low relative shares of inconsistent initial commitments that result in consistent output commitments Conversely, the grey band connecting inconsistent initial and output commitment at the bottom of these plots dominates. This corresponds to the inconsistency preserving cases (IP), which make up most of all cases of inconsistent initial commitments.

A horizontal comparison in Figure 10.2 between global optima and fixed points for the same configuration reveals that there are almost no noteworthy differences. Only the consistency eliminating cases (CE, the band leading from consistent initial commitments to inconsistent outputs) are slightly more pronounced for global optima than fixed points for the second configuration of weights.

Discussion For the first configuration of weights, that is, $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$, the formal exhibits substantive revisionary power with respect to inconsistent initial commitments. Inconsistencies from the initial commitments are eliminated quite successfully. In this case, the formal model escapes inconsistency preservation as an aspect of (Conservativity).

Moreover, the model performs very well with respect to the other cases, which are less relevant to the conservativity, but interesting from the viewpoint of model validation. The model mostly preserves initial consistency (CP), and hence there are very few cases of consistent initial commitments that result in inconsistent output commitments (CE). For plots that convey a more clear picture of the relative share of consistency cases than diagrams with overlapping bands, see Appendix D.1.2).

[6] Note that the small differences in relative shares of consistent initial commitments are due to the model producing different numbers of global optima and fixed points for the same initial commitments.

The situation changes dramatically for the second configuration, which gives a lot of weight to faithfulness $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$. The formal model is extremely conservative with respect to inconsistency preservation. In this case, "garbage in - garbage out" applies. The vast majority of inconsistencies from initial commitments are preserved in output commitments of RE outputs.

I suppose that the salient difference between configuration can be explained in terms of different trade-offs between account and faithfulness. Note that this line of thought applies to global optimisation as well as semi-global optimisation during commitment adjustment steps in a process of equilibration. The formal model does not enforce dialectical consistency in commitments, but recommends it through the achievement function. Dialectically consistent theories should account for commitments. This indirectly exerts pressure to render the commitments consistent as well. Otherwise, there are penalties for contradictions in the measure for account $A(C, T)$. The amount of pressure to revise commitments is increased if we give more weight to account than faithfulness ($\alpha_A > \alpha_F$). The weight for systematicity α_S plays a coordinate role, as it restrains the weight for faithfulness in the boundary condition $\alpha_A + \alpha_S + \alpha_F = 1$ in cooperation with α_A. This is underwritten by the results from additional configurations in Appendix D.1.1.

The streamlining baseline is not included because it reaches consistent outputs in every case. Now, we have seen that there are cases in which the model produces inconsistent output commitments. Does this render the model worse than the baseline? I do not think so. The formal model has an advantage over mere streamlining independent of any study findings. The procedure that streamlines the starting point is "rigged" to achieve consistency (as well as full and exclusive account) at all costs. Consistency is established by fiat. In contrast, the model recommends consistent commitments. Only in case of a positive trade-off with other virtues, consistency is established. I count the default model's flexibility with respect to consistency as a strength. Furthermore, the present findings back up the model. Without being "rigged ", it still produces consistent outputs in a wide range of cases for specific configurations of weights.

10.5 Does RE Make Views More Systematic?

Background Typically, an agent enters RE with more or less unsystematic commitments about a subject matter. Elgin (1996, 102), for example, speaks

of initial commitments as a "motley crew". Proponents as well as critics of RE regularly mention the methodological instruction that an agent should systematise their initial commitments. However, systematisation is not spelled out to play any active role in the process of equilibration or the characterisation of an equilibrium state. In contrast, more recent and elaborate accounts of RE assign an active and specific role to systematicity. Systematisation of commitments is achieved through a virtuous theory that accounts for the commitments.

If RE falls prey to (Conservativity), we should expect that RE is "no more than a re-shuffling of one's initial prejudices" (Brandt, 1985, 7), and hence result in little progress concerning systematisation.

Method In order to investigate whether there is progress in systematisation between initial and output commitments we need a measure of systematicity that applies to them.

The default model offers a measure of a theory's systematicity $S(T)$, but it is biased towards theories that contain a single element (see Section 8.1), and it does not translate well to a measure of commitment systematicity. Instead, and well suited for the deductive setting of the formal framework, the degree of how simply the commitments can be axiomatised by themselves is a good indicator of how well a position can be systematised by a theory. If the commitments are inferrable from a small subset (their axiomatic base), a lot of inferential relations from the dialectical structure in the background can be exploited to systematise the commitments.

I implement *axiomatic systematicity* through minimal axiomatic bases (see Section 10.3). In order to see how simply a position P can be axiomatised by its own elements, we can take the axiomatic base from the source $\mathcal{S}$ containing P and all of its subsets. If a position cannot exploit the inferential relations given by the arguments in the dialectical structure in the background, the axiomatic base of P has to include many elements. In the worst case, the axiomatic base of P is P itself.

The size of a minimal axiomatic base for a position P taken from a source $\mathfrak{S}$ that is restricted to the position and its subsets does not yet make a good measure for axiomatic systematicity. It is in need of normalisation because otherwise, the strategy of removing sentences from a position would result in an increase of axiomatic systematicity.

Thus, I propose to study the following, normalised measure of axiomatic

systematicity. Let C be a dialectically consistent position and let $mab(C)$ denote a minimal axiomatic base of C from the source of subsets of C (including C itself). $mab(C)$ may not be uniquely determined, but due to minimality, $|mab(C)|$ is unique. The axiomatic systematicity of C is defined as

$$S_{axiom}(C) = 1 - \frac{|mab(C)|}{|C|}$$

The measure returns the minimal value of 0 if and only if the minimal axiomatic base of C is just C itself. In this case, no non-trivial inferential relation in the dialectical structure can be exploited to axiomatise C, and the position C is maximally unsystematic. In reverse, the measure is maximal if and only if a position with a single sentence axiomatises a complete position C. Thus, the maximal value depends on the size of the sentence pool, and it approaches 1.0 for larger sizes.

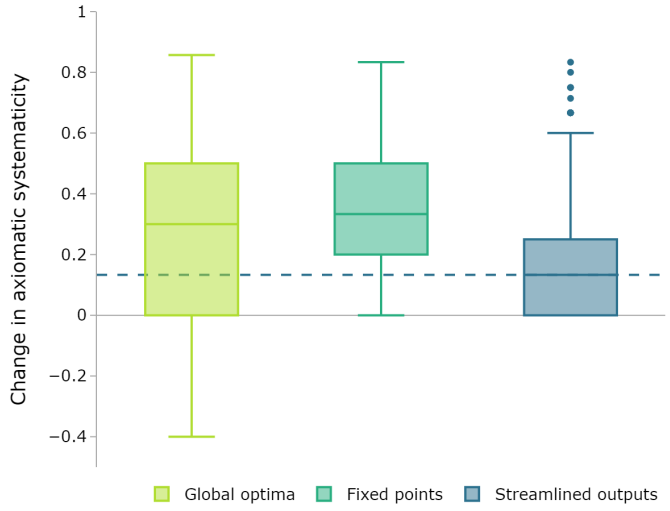
Take, for example, the initial commitments $C_0 = \{3, 4, 5\}$ from the standard example (see Section 7.1.2). Given the dialectical structure of the standard example, there are no inferential relations among the elements of C_0 that could be exploited to construct a minimal axiomatic base. Consequently, the minimal axiomatic base of C_0 amounts to C_0 itself, which shows that the initial commitments are quite unsystematic, and indeed, $S_{axiom}(C_0) = 0$.

For the standard configuration of weights, the set of commitments in the global optimum (fixed point) reached from C_0 is $C = \{1, -2, 3, 4, 5, -6\}$. Now, the minimal axiomatic base of C is $\{1\}$ that exploits many inferential relations. Here, $S_{axiom}(C) = \frac{5}{6}$ which is a good indication of more systematisation among the output commitments. The difference $S_{axiom}(C) - S_{axiom}(C_0)$ is positive for the present example. However, this difference can be negative if the output commitments are less axiomatically systematic than the input commitments.

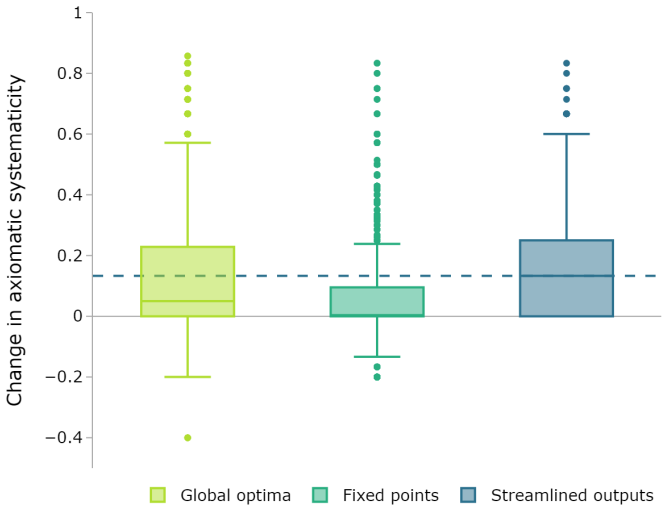
The change of axiomatic systematicity between initial and output commitments reached from this starting point is the endpoint for the present study.

I limit the study to simulations with consistent initial and output commitments that have an axiomatic base. Of course, it would also be possible to construct an axiomatic base from a maximally consistent subposition of a inconsistent position. However, the streamlining procedure implements this exact mechanism for inconsistent initial commitments. Thus, including axiomatic systematicity for inconsistent initial commitments is not very interesting or might even skew the results in favour of the streamlining procedure.

If RE is conservative in not exerting enough pressure to systematise one's views, we should expect to see little progress in axiomatic systematicity between initial and output commitments.



(A) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$



(B) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$

FIGURE 10.3: Change in axiomatic systematicity between initial and output commitments. The dashed, blue baseline depicts the median change between initial and streamlined commitments. Positive (negative) values indicate progress (regress) in systematisation.

Results Figure 10.3 summarises the results for the difference in axiomatic systematicity between initial and output commitments. The solid horizontal line centred at 0 corresponds to no change in axiomatic systematicity between initial and output commitments. The position of the boxes thus reveal that the outputs achieve some progress, on average. Global optima and fixed points occasionally result in a negative change meaning that the output commitments are less axiomatically systematic than the initial commitments.

The streamlining procedure (median: 0.13, IQR: 0.00–0.25) provides the conservative baseline (blue, dashed line). Being higher than the baseline means more progress in terms of axiomatic systematicity.

For the first configuration in (A) in Figure 10.3, global optima (median: 0.30, IQR: 0.00–0.50) and fixed points (median: 0.33, IQR: 0.20–0.50) exceed the streamlining baseline, on average. In case (B), both global optima (median: 0.05, IQR: 0.00–0.23) and fixed points (median: 0.00, IQR: 0.00–0.10) fall short of the streamlining baseline and achieve only little progress.

Discussion For the first configuration of weights, that is, $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$, which puts a lot of weight to account, the formal model escapes what conservativity would lead us to expect. The formal model exerts enough pressure to systematise the initial commitments, resulting in outputs that can be axiomatised more simply than their inputs or streamlined outputs.

The streamlining baseline procedure works with a deliberately small set of candidates (the sub-positions of the initial commitments). In contrast, global and semi-global optimisation have much more options at hand. Thus, it is quite remarkable that the formal model does not exploit this potential for the second configuration of weights $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$. There is very little progress in systematisation in terms of axiomatic systematicity

Again, we can observe a positive impact of theoretical virtues (systematicity and account) on axiomatic systematicity, where they have more weight than faithfulness (see Appendix D.1.1). They join forces as follows: Systematicity recommends theories that strike a good balance between simplicity and scope, or in other words, positions that axiomatise well with few resources. Account advocates for a good fit between current commitments and the theory. Optimally, the output theory is a minimal axiomatic base of the output commitments. Otherwise, there are penalties for systematicity or account.

10.6 Conclusion

The results found in the ensemble study present an overall consistent picture. The model performs better than the baseline with respect to all operationalised aspects of conservativity for the standard as well as additional configurations of weights. The latter configurations of weights help to see the bigger picture (see also Appendix D.1.1): In general, the formal model of RE performs better than the conservative baseline if account receives more weight than faithfulness ($\alpha_A > \alpha_F$). Conversely, the model is equally as conservative, or even more conservative than the baseline if faithfulness is given higher weight than account ($\alpha_F > \alpha_A$). The weight for systematicity α_S has also a positive effect on non-conservative behaviour, but to a lesser extent than account. This may be due to the fact that systematicity and account jointly keep faithfulness at bay due to the boundary condition $\alpha_A + \alpha_S + \alpha_F = 1$.

So, it is plausible that the weight for faithfulness α_F is directly linked to the conservativity of the default model in all operationalised aspects of conservativity. This is to be expected, if we take the measure for faithfulness into consideration. $F(C \mid C_0)$ involves a Hamming distance between initial and current commitments. The higher this distance, the lower the value of $F(C \mid C_0)$. Consequently, high weightings for α_F put pressure on the model to minimise the Hamming distance between initial and current commitments resulting in the preservation of the input. This is apparent for the aspect of minimal change operationalised with the edit distance (Section 10.3). Moreover, the conservative impact of faithfulness surfaces for other aspect, inconsistency preservation, and axiomatic systematicity, as well, because they emerge from changes on the level of sentences.

If faithfulness controls the default models conservativity, which turns out to be problematic for RE, why should we not abandon it all together? Interestingly, faithfulness cannot be removed from the achievement just like that. Proposition 3 in Chapter 8.1 shows that $\alpha_F = 0$ is a bad idea, as it yields exactly singleton theories and their dialectical closure as global optima.[7]

Note that the conservative impact of faithfulness may stem from the rather drastically operationalisation in the default model. This is at best a very crude implementation of the idea of "respecting the input" (Baumberger and

[7]For fixed points, the question, whether setting α_F to 0 produces plausible outcomes, remains open. The fact that RE processes start by selecting a theory in view of the initial commitments may be enough of a "tie" to the starting point. Alternatively, RE processes could be equipped with a procedural variant of faithfulness that exhibits the Markov property, that is its measure depends on current state only.

Brun, 2021), which is about not giving up commitments without good reason instead of not moving too far beyond the starting point.

The finding that we can set the parameters of the default model to exhibit overly conservative, as well as revisionary behaviour illustrates the default model's flexibility to capture very different strategies. Moreover, we can turn this finding into a conciliatory point for the debate about RE in general. Detractors of RE, who accuse the method of RE being to conservative, are not arguing beside the point. If we construe RE to involve the idea that an agent is faithful to their starting point to some degree, conservativity is a non-negligible problem that has to be taken seriously by proponents of RE.

However, the results also suggest a solution how RE escapes the threat of (Conservativity). There is a range of weight configurations for which the model has sufficient revisionary power to block overly conservative strategies. In particular, the favourable configurations give more weight on account and systematicity, which implement theoretical virtues in the model. Thus, in general, including additional theoretical virtues into RE may have a positive impact on performance with respect to conservativity. This is apparent for the default model in comparison to the baselines of streamlining and simplistic adaptation. The baselines are almost devoid of theoretical virtues, only consistency and fit is enforced by streamlining without allowing for trade-offs.

I take the findings of this study to underwrite the following point. Critics that object to RE on the basis of (Conservativity) under-appreciate the role of systematisation. Spelling out systematisation in terms of theoretical virtues and assigning it an active role during equilibration processes or for the characterisation of equilibrium states makes RE stronger. Systematisation provides more incentive to revise initial commitments than merely escaping incoherence with minimal adjustments.

The fact that the model performs better than conservative baselines for a specific range of weight configurations provides some *ex post* justification to those configurations. Outputs resulting from non-conservative weight configurations seem to be better justifiable from a coherentist perspective (consistent, more systematic and well-accounted outputs) without giving up on the weakly foundationalist spin on RE that is implemented by faithfulness in the formal model. We can back up this point further if other objections against RE can be resolved for the same range of weight configurations. I

take up this task in the next chapter.

Appendix

D.1 Robustness

D.1.1 Configurations of Weights

Apart from the two paradigmatic configurations of weights used to present the results of the ensemble study, there are five additional configurations spread out over the parameter space (see Figure 9.8). Combined with the result that sets of configurations that yield a global optimum form convex (and hence, connected) regions, this should give us good coverage of what is going on for a wide range of configurations. The results are collected in tables for edit distances (Table D.1), inconsistency preservation (Table D.2), and axiomatic systematicity (Table D.3).

configuration	global optima	fixed points	streamlined outputs
(0.35, 0.55, 0.10)	3 (2–4)	2 (2–3)	2 (1–2)
(0.55, 0.35, 0.10)	3 (3–4)	3 (2–4)	2 (1–2)
(0.70, 0.20, 0.10)	3 (2–4)	3 (2–3)	2 (1–2)
(0.55, 0.20, 0.25)	2 (2–3)	2 (2–3)	2 (1–2)
(0.46, 0.10, 0.44)	2 (2–3)	2 (1–3)	2 (1–2)
(0.10, 0.55, 0.35)	1 (1–2)	1 (1–1)	2 (1–2)
(0.10, 0.35, 0.55)	1 (0–2)	1 (0–1)	2 (1–2)

TABLE D.1: Median (IQR) edit distance between initial and various output commitments grouped by configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$.

The most important insights are the following. For configurations that put very low weight on faithfulness ($\alpha_F = 0.10$), the formal model performs robustly better than the streamlining baseline (edit distance, axiomatic systematicity), and exhibits desirably low shares of inconsistency preserving cases. Note that for the three best performing configuration, it seems that giving more weight to account than systematicity improves the results concerning even further. Next, configurations that give more weight to faithfulness

configuration	global optima	fixed points
(0.35, 0.55, 0.10)	0.21	0.14
(0.55, 0.35, 0.10)	0.05	0.01
(0.70, 0.20, 0.10)	0.00	0.00
(0.55, 0.20, 0.25)	0.54	0.26
(0.46, 0.10, 0.44)	0.40	0.24
(0.10, 0.55, 0.35)	0.72	0.72
(0.10, 0.35, 0.55)	0.91	0.90

TABLE D.2: Relative share of inconsistency preserving output commitments among cases with inconsistent initial commitments grouped by configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$.

configuration	global optima	fixed points	streamlined outputs
(0.35, 0.55, 0.10)	0.13 (0.00–0.33)	0.17 (0.00–0.40)	0.13 (0.00–0.25)
(0.55, 0.35, 0.10)	0.30 (0.00–0.50)	0.33 (0.20–0.50)	0.13 (0.00–0.25)
(0.70, 0.20, 0.10)	0.42 (0.21–0.57)	0.30 (0.17–0.50)	0.13 (0.00–0.25)
(0.55, 0.20, 0.25)	0.30 (0.17–0.50)	0.30 (0.17–0.46)	0.13 (0.00–0.25)
(0.46, 0.10, 0.44)	0.32 (0.17–0.50)	0.26 (0.14–0.43)	0.13 (0.00–0.25)
(0.10, 0.55, 0.35)	0.03 (0.00–0.20)	0.00 (0.00–0.05)	0.13 (0.00–0.25)
(0.10, 0.35, 0.55)	0.05 (0.00–0.23)	0.00 (0.00–0.10)	0.13 (0.00–0.25)

TABLE D.3: Median (IQR) change of axiomatic systematicity between initial and output commitments grouped by configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$.

than systematicity but less than account, $(0.55, 0.20, 0.25)$ and $(0.46, 0.10, 0.44)$, the model yields intermediate results that fall between the best performing configurations and the streamlining baseline. This speaks in favour of systematicity having a positive effect on the revisionary power of the formal model. Finally, if faithfulness receives more weight than account, i.e., for the configurations $(0.10, 0.55, 0.35)$ and $(0.10, 0.35, 0.55)$, the model performs as bad as, or even worse than the streamlining baseline.

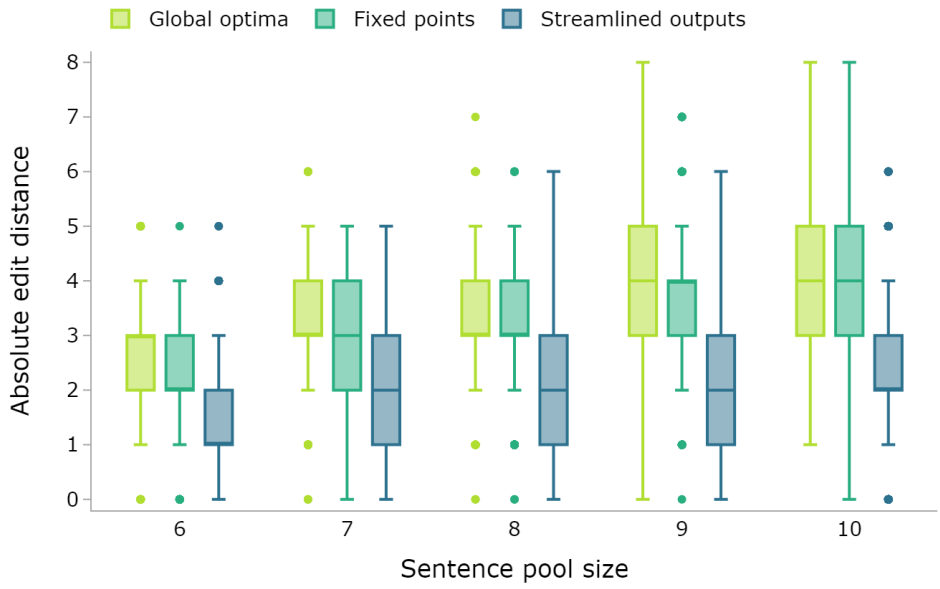
There is a noteworthy observation concerning the standard configuration used in (Beisbart, Betz, and Brun, 2021) for axiomatic systematicity in Table D.3. For $(\alpha_A, \alpha_S, \alpha_F) = (0.35, 0.55, 0.10)$, global optima (median: 0.13, IQR: 0.00–0.33) perform roughly as the streamlining baseline (median: 0.13, IQR: 0.00–0.25) with respect to the change in axiomatic systematicity between initial and output commitments. I suspect that this stems from increased occurrence of "trivial" outputs, i.e., states that consist of a theory with and a set of commitments both containing exactly one element. Singleton commitments can only be axiomatised by themselves, which leads to the minimal value of 0 for axiomatic systematicity. The production of such "trivial" is likely due to the high weight for systematicity combined with the model's bias towards singleton theories (Section 8.1).

D.1.2 Sentence Pool Sizes

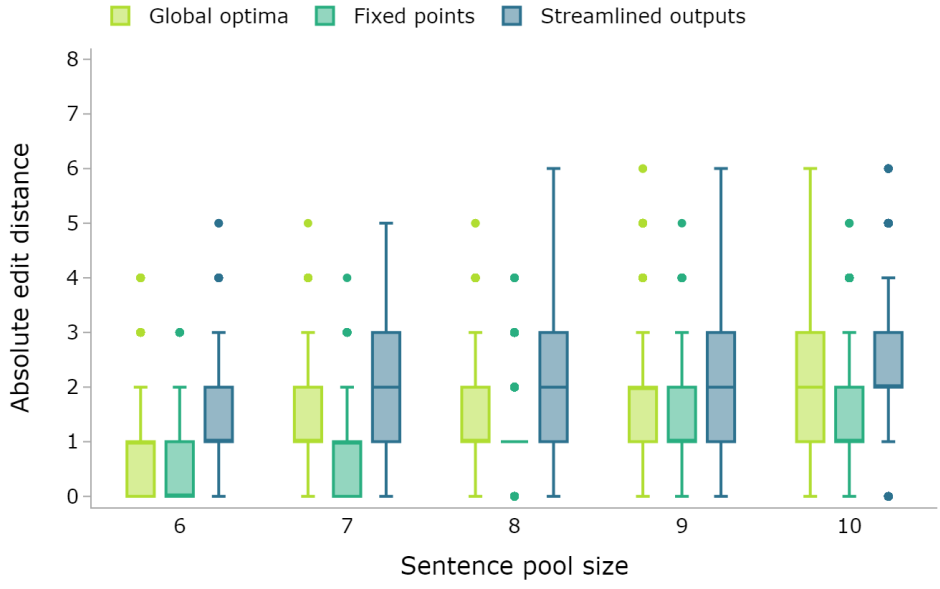
An additional ensemble of simulations serves to study the influence of the sentence pool size on conservativity. It consists of the following simulation setups: The sentence pool sizes range from 6 to 10 unnegated elements, there are 50 randomly generated dialectical structures per sentence pool size, and 10 initial commitments per dialectical structures. This is combined with (A), the default configuration $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$, as well as (B), the conservative configuration $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$, this results in a total of 5,000 simulation setups.

The grouped results are depicted for edit distances (Figure D.1), consistency cases (Figure D.2 and Figure D.3), and axiomatic systematicity (Figure D.4).

In the box plots for edit distances, the effect of increasing the sentence pool size is directly visible. Medians and inter quartile ranges (boxes) increase with higher sentence pool sizes. However, this applies to all kinds of outputs, and in such a way, that does not alter the standing of RE outputs and the streamlining baseline. For the first configuration (A), the outputs of

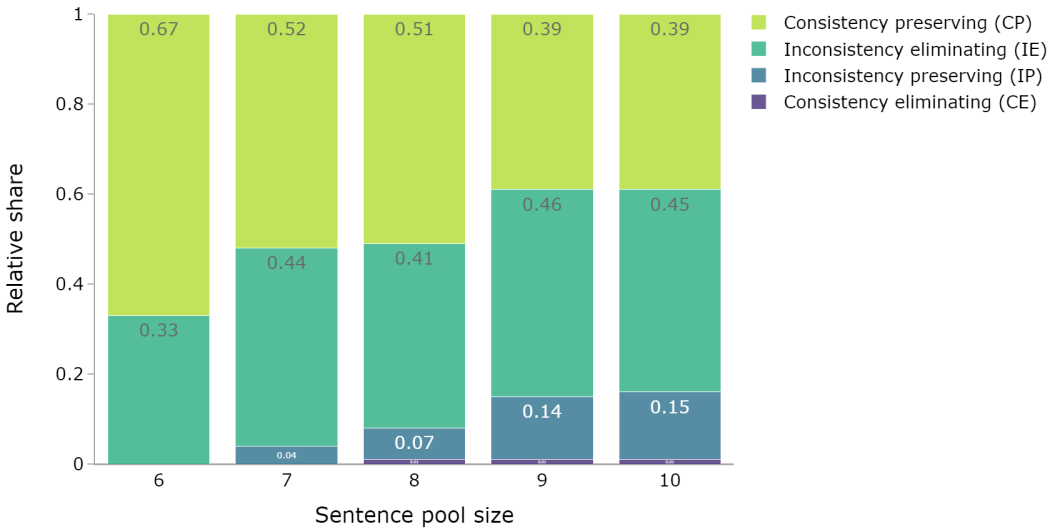


(A) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$

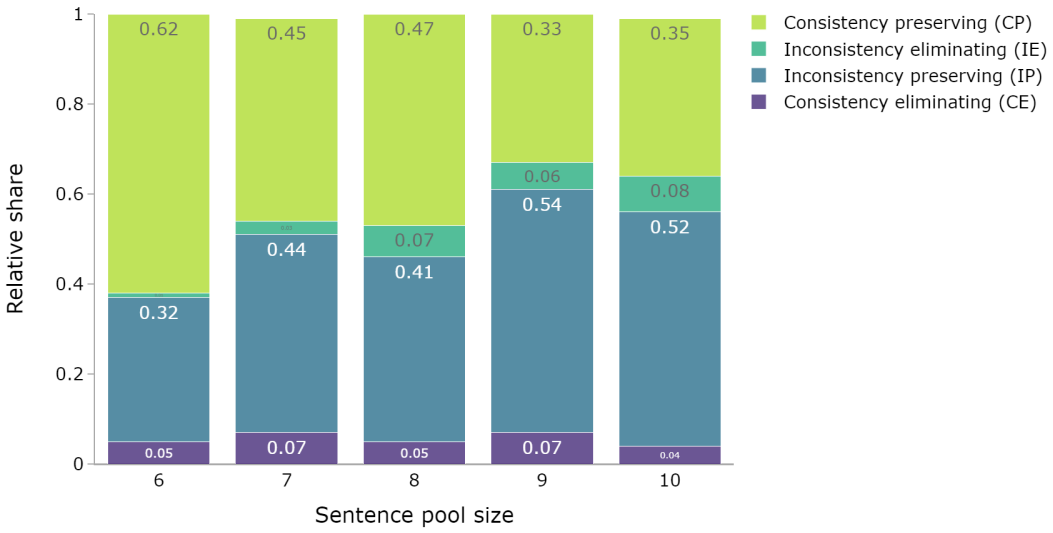


(B) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$

FIGURE D.1: Edit distance between initial and output commitments grouped by sentence pool size.

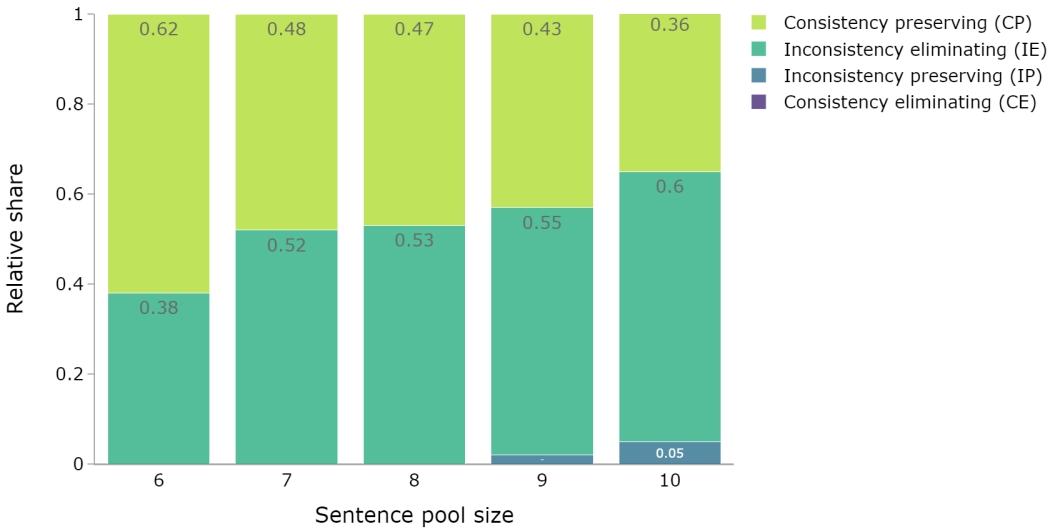


(A) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$

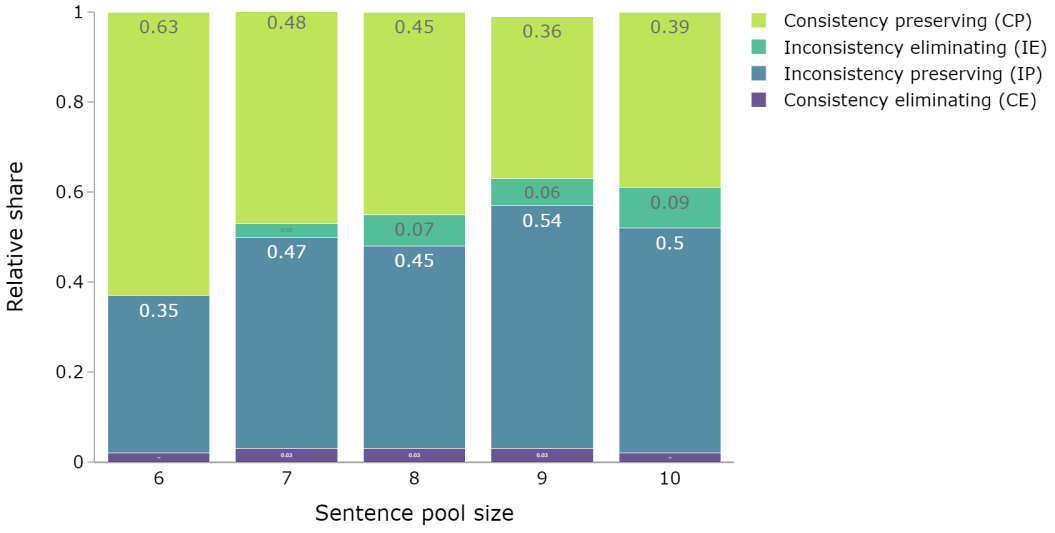


(B) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$

FIGURE D.2: Relative share of consistency cases for global optima grouped by sentence pool size.

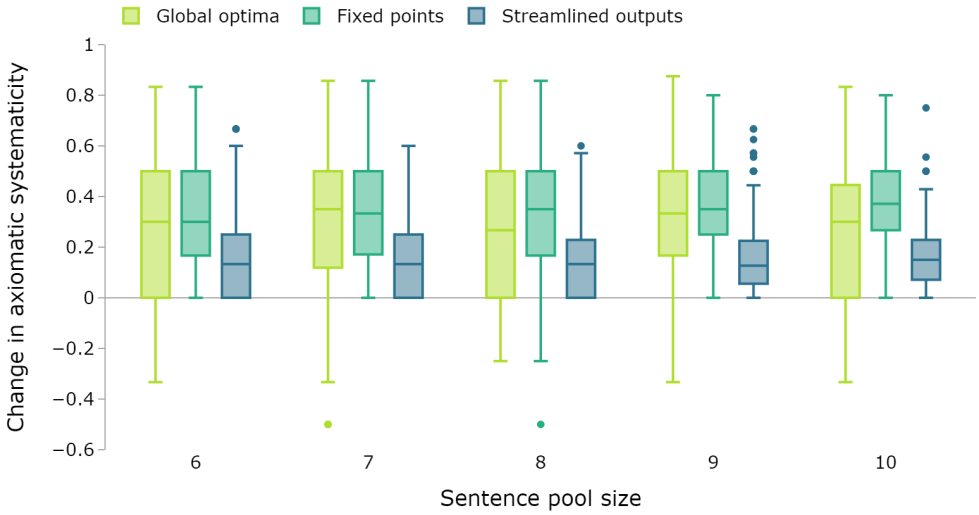


(A) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$

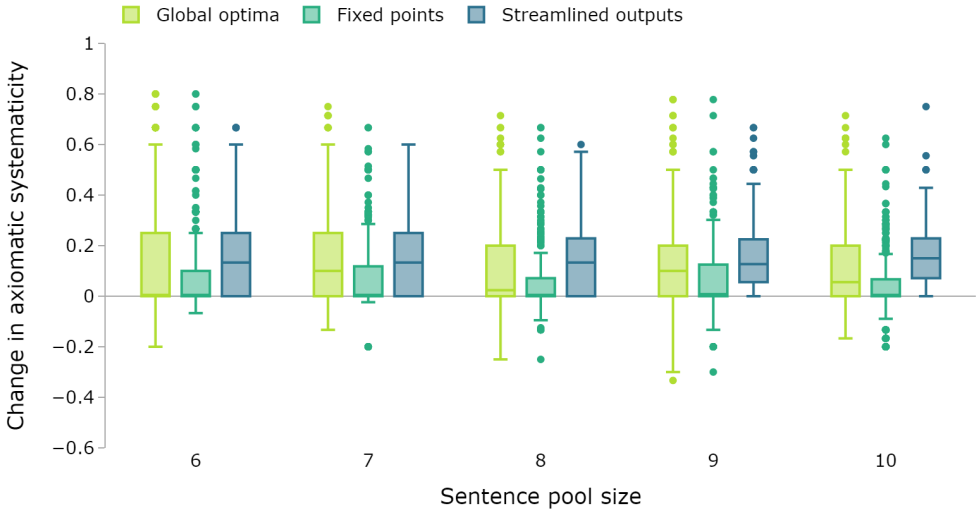


(B) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$

FIGURE D.3: Relative share of consistency cases for fixed points grouped by sentence pool size.



(A) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$



(B) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$

FIGURE D.4: Change in axiomatic systematicity between initial and output commitments grouped by sentence pool size.

the formal model of RE perform better than the streamlining baseline. In contrast, the RE outputs do not perform better, or even worse than the streamlining baseline for all sentence pool sizes and configuration (B). Hence, the results are robust with respect to the small variations in the sentence pool size presented here.

For axiomatic systematicity, there is no notable effect of the sentence pool size to the results, which are also consistent with the findings that I presented before.

Consistency cases take mostly the same line except that there is a notable increase of inconsistency preserving cases for fixed points in Figure D.3, (A), which is not that pronounced for global optima in Figure D.2, (A). I cannot offer a full explanation, but I conjecture that this is due to the procedural tie to initial commitments during equilibration, which is absent from global optimisation.

Chapter 11

Does Reflective Equilibrium Help Us Converge?

Recall the objection of no-convergence to RE that I introduced informally in Chapter 3:

(No-Convergence) RE is not able to achieve converging, non-substantially-disagreeing outputs.

(No-convergence) is presented as a problem for the justificatory power of RE for various reasons covered in Section 3.3. They may be summarised by the worry of Kelly and McGrath (2010) that RE is too *weak* as an account of justification. In Section 3.3, I identified three aspects of convergence that surface in the literature on RE. We can frame them as questions: Does RE yield a unique output? Does RE promote agreement? Does RE allow for "anything" goes?

The aim of this chapter is to address the no-convergence objection and provide answer to these questions on the basis of simulations run by the computer implementation of the formal model of RE. This puts us again in a position to study the influence of theoretical virtues in RE, and in particular, the configuration of weights.

The work is organised as follows: In Section 11.1, I provide information about the simulations. In sections 11.2, 11.3 and 11.4, I separately address the three aspects of convergence under the rubric of method-results-discussion. Finally, I draw lessons for the informal debate about RE in Section 11.5. I relegate a robustness analysis of the presented results to Appendix E.1.

11.1 Information about Simulations

It is desirable to have a sample including many different simulation setups, but computational limitations require a trade-off between the number of dialectical structures, the number of initial commitments, and the number of configurations of weights. I use the seven configurations of weights selected in Section 9.3, and present results for the same paradigmatic configurations as in the previous chapter, namely

$$(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.1) \text{ and } (\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55).$$

The results of additional configurations of weights are available in Appendix E.1.1.

The "two-point ensemble" serves to compare pairs of equilibration processes in Section 11.2 and Section 11.3. It comprises 13,000 randomly generated dialectical structures[1], but only two random sets of initial commitments per structure.[2] This results in a total of 182,000 simulation setups.

The "full-spectrum-ensemble" is designed to investigate the allegation of "anything goes" (Section 11.4), which often takes off from the assumption that there are many and drastically different inputs. The idea of maximally diverse initial commitments can be operationalised by setting up RE simulations from the *full spectrum* of initial commitments, which consists of all non-empty and minimally consistent positions. For example, there are $3^7 - 1 = 2{,}186$ non-empty positions that can serve as initial commitments for a sentence pool size of 7. In order to accommodate this high number of initial commitments, the second ensemble includes fewer randomly generated structures (30). The full-spectrum-ensemble thus comprises 459,060 simulation setups.[3]

For every simulation setup, all fixed points reached due to branching processes and all global optima have been collected. I also include the streamlining procedure (introduced in Section 3.2 and operationalised in Section

[1] Dialectical structures are created randomly with a sentence pool of size 7, and 5–8 arguments consisting of 1–2 premises and a conclusion. It is ensured that the resulting dialectical structure is minimally orderly. The arguments are jointly satisfiable, they are not question-begging by repeating the conclusion among premises, they avoid flat contradictions, and they do not use the same premises for different conclusions.

[2] The initial commitments are also randomly chosen, but it is ensured that the number of sentences, about which positions of a pair differ, are spread out evenly instead of being normally distributed.

[3] Raw data and exploratory notebooks including interactive plots are retrievable from https://github.com/free-flux/virtuously-circular/tree/main/chapter-11.

10.2), and use it once again as a baseline for comparison. The streamlining procedure is conservative by design, and this is corroborated by the results from the previous chapter. Now, conservativity fuels the no-convergence objection: If there is not enough incentive to revise inputs and the inputs differ from each other, then the differences are likely to be preserved in the outputs. Consequently, the streamlining procedure is also an interesting point of reference to asses the model's performance with respect to aspects of convergence.

As in the previous chapter, I focus on sets of output commitments and bracket the theories of outputs. Note that theories are frequently included in the set of outputs commitments anyway. In the two-point-ensemble, 91.9% of all global optima, and 91.4% of all fixed points commitments contain the theory as a subset. For the full-spectrum-ensemble the respective results are 93.9% (global optima) and 93.7% (fixed points).

11.2 Does Reflective Equilibrium Yield a Unique Output?

Methods Kelly and McGrath (2010, 337) distinguish between *intrapersonal* and *interpersonal* convergence, i.e., whether i) an individual agent with a single starting point, or ii) a group of agents with different starting points reach a unique output, respectively.

Intrapersonal convergence to a unique output can be tracked easily in the formal model and its computer implementation. As it stands, the formal model does not implement interactions between agents. Epistemic states of other agents are not taken into consideration at any point in an equilibration process or for the evaluation of epistemic states.

Intrapersonal convergence might not obtain in the formal model for the following reason: Even if the model is provided with a dialectical structure, a configuration of weights and some initial commitments, some adjustments during the process of equilibration may be underdetermined. There may be multiple candidates in an adjustment step that perform equally well according to the achievement function. By design, such ties are resolved with random choices that cause an equilibration process to branch out. If we track every branch of an equilibration process, we can examine whether they lead to different fixed points.[4]

[4]Note that the underdetermination of adjustments is the only form of path-dependency in the formal model. An equilibration process proceeds by semi-global optimisation, i.e.,

Similarly, multiple global optima can arise from ties within the achievement function. Consequently, the model might produce multiple fixed points and global optima from a single simulation setup. In this case, the formal model would not exhibit intrapersonal convergence to a unique output.

Interpersonal convergence of two agents can be studied by considering the pairs of simulation setups in the two-point-ensemble (two sets of initial commitments in the same dialectical structure). If both simulation setups exhibit intrapersonal convergence to a unique output, do they reach the same output?

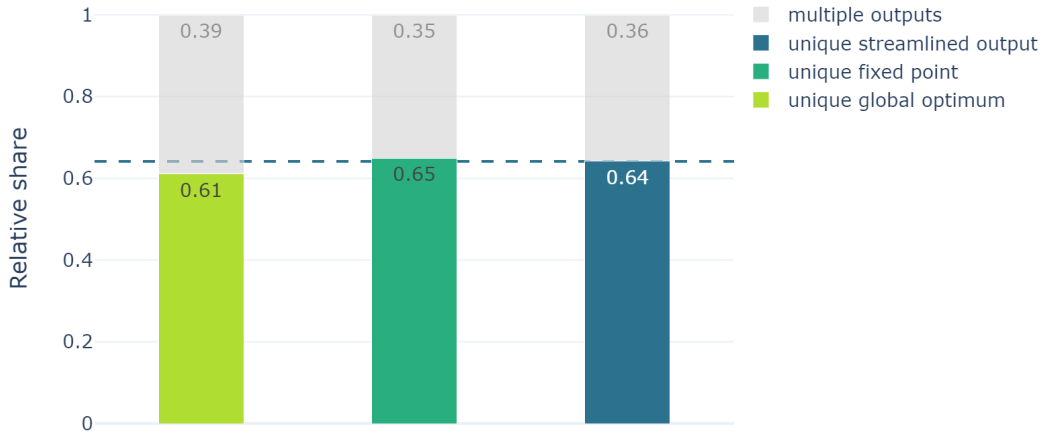
Results Concerning intrapersonal convergence, Figure 11.1 depicts the relative share of simulation setups that yield a unique output. There are no notable differences between the performances of global optima, fixed points and streamlined outputs, or between the configuration of weights. In roughly two thirds of cases, a simulation setup results in a unique output.

For interpersonal convergence, pairs of simulation setups have been restricted to those that both reach a unique output. Otherwise, a unique output is not feasible. Subsequently, the relative share of simulation setups that reach the same unique output has been determined. The results are displayed in Figure 11.2. Overall, it is notable that the relative shares are rather low, and especially, in comparison to the results for intrapersonal convergence in Figure 11.1.

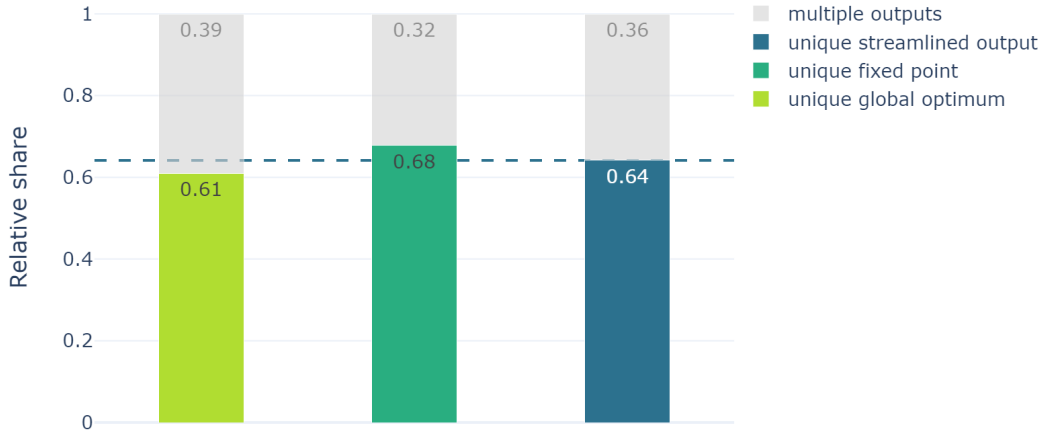
For the first configuration in Figure 11.2 (A), global optima and fixed points perform substantially better than the streamlining baseline. Global optima slightly surpass fixed points. For the second configuration in (B), they fall short of the baseline, and produce unique outputs from paired simulation setups very rarely.

Discussion The formal model achieves intrapersonal convergence in relatively many cases. In these cases, the formal model restricts adjustments and states so as to narrow down the range of options to a unique output. The observation that there is a configuration of weights that leads to the occurrence of interpersonal convergence to a unique output from a pair of simulation setups in roughly a fifth of cases is respectable given the random generation of dialectical structures and sets of initial commitments.

the evaluation of all theory/commitments candidates in a theory/commitments adjustment step. There are no series of adjustments of individual elements in a position whose order could become relevant.

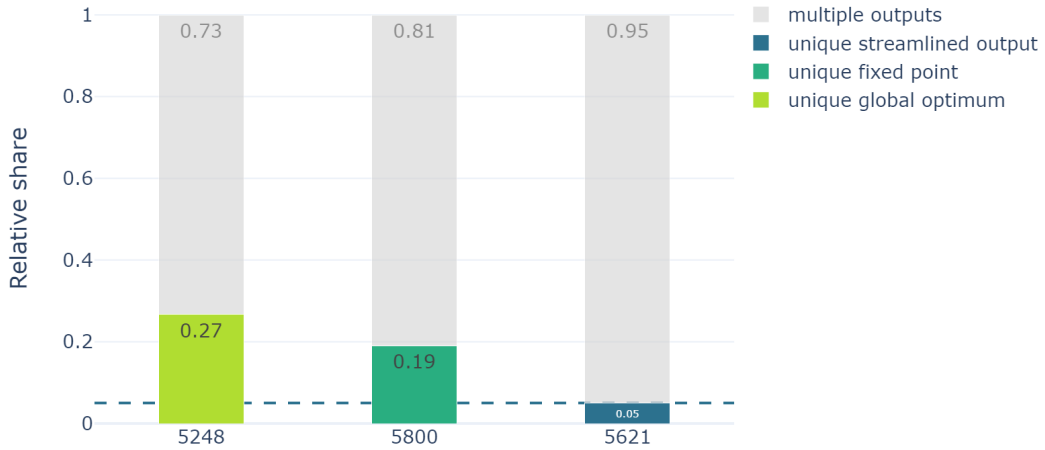


(A) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$

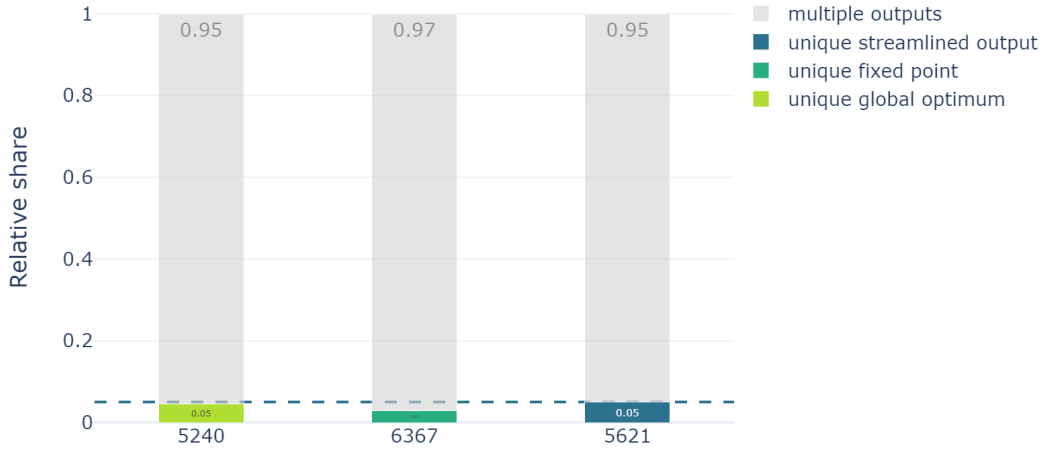


(B) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$

FIGURE 11.1: Relative share of simulation setups that result in a unique output. The total number of simulation setups per configuration is 26,000.



(A) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$



(B) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$

FIGURE 11.2: Relative share of paired simulation setups that result in the same, unique output. The number of paired simulation setups that both yield a unique output is indicated below the bar.

Concerning intrapersonal convergence, the model performs roughly as good as the streamlining baseline. Is this problematic for the evaluation with respect to convergence? Recall that the formal model allows for different trade-offs in contrast to the streamlining procedure, which is much more rigid by design and does not optimise (semi-)globally. In view of this difference, the formal model catching up with the streamlining baseline with respect to intrapersonal convergence is a satisfactory result.

However, there are cases where intrapersonal and, even more so, interpersonal convergence do not obtain. Moreover, the comparison of individual to pairs of simulation setups reveals a substantive decrease in the relative share of (pairs) of simulation setups that converge to a unique output. Presumably, the model preserves some differences between the paired sets of initial commitments throughout the process of equilibration or in global optimisation. One has to expect that considering more than two simulation setups would further erode the prospects of achieving interpersonal convergence to a unique output.

Of course, one could try to impose additional constraints in an attempt to reach unique outputs. For example, one could lower the weight for faithfulness even further or require substantial overlap in pairs of sets of initial commitments. However, I doubt whether such attempts could keep up with the intricacies of de-idealised, informal RE settings. First of all, realistic examples would contain more than seven sentences. Take for example Rechnitzer's detailed application of RE to the justification of a precautionary principle (Rechnitzer, 2022), which involves easily more than one hundred elements. Transferring these elements to sentences in the present formal framework would yield more than 3^{100} (515 septilliard) positions. Apart from not being computationally feasible, such an example bears exponentially more potential for ties, and hence might result in much more outputs. Moreover, in an informal setting, there are no ready-made numerical measures to evaluate epistemic states according to RE desiderata or straightforward solutions to handle trade-offs. Such complications might also contribute further to the multiplication of results.

At some point, it becomes doubtful whether the constraints needed to ensure uniqueness would yield even remotely plausible constraints that would be insightful for informal applications of RE. Given the present results in a highly simplified and idealised formal model of RE, the hopes are very dim that RE in an informal setting could do better.

A more promising move is to admit that uniqueness is too stringent as

a condition for convergence on RE. Uniqueness demands complete coincidence among outputs. This blocks the view of more subtle forms of agreement. As a consequence, the failure to produce a unique output gives us motivation to adopt a pluralist stance on justification with RE, as some authors already do, e.g., (Elgin, 1996, 135) or (Rechnitzer, 2022, 236).

11.3 Does RE Promote Agreement?

Methods "Agreement" and its cognates remain rather vague in the literature. There is a notable exception, however, which offers a fruitful starting point for formalisation: Tersman (1993) distinguishes between two "systems" of beliefs being *incompatible* and *differing* from each other. According to him, two systems A and B are *incompatible* if A contains an element p such that there are elements in B that jointly imply that p is false (Tersman, 1993, 84). In contrast, two systems A and B *differ* if A contains an element that is not in B, or vice versa (Tersman, 1993, 105). As A and B may differ with respect to more or less elements, difference becomes a gradual notion.

Tersman's treatment of incompatibility and differences translates very well to the framework of the formal model. Compatibility amounts to the requirement that positions are consistent with each other given the arguments of the dialectical structure. Given a dialectical structure, two positions are *dialectically compatible* if and only if their set-theoretic union is dialectically consistent. In such cases, agents could aggregate their individual outputs of RE, e.g., by taking the union of their commitments, without running into contradictions. So construed, compatibility is a categorical feature of positions. It does not take the number or the severity of conflicts into consideration.

We can complement compatibility by a gradual notion of *similarity* between positions on the more fine-grained level of sentences. The following generalises the measure of *normalised agreement* of Betz (2012, 39) to partial positions.[5] I rename the measure *(normalised) similarity* to prevent confusion with my present use of "agreement": (normalised) similarity operationalises agreement on the propositional level.

Let n be the size of the sentence pool, and assume that P and Q are minimally consistent positions. Then,

$$\Delta(P, Q) = \frac{D_{0,1,1,2}(P, Q)}{2n}$$

[5] Betz's measure applies to complete positions exclusively. My proposal is a natural extension as the measures are identical for complete positions.

is the *normalised Hamming distance* between P and Q that measures differences by summing penalties for expansion, contractions and contradictions over all sentences. For the definition of D, see Section 7.1.2. Note that the penalties of the Hamming distance $D_{0,1,1,2}$ are chosen such that contradictions receive a penalty of 2, while it penalises contractions and expansions with 1. This choice reflects the idea that two positions differ to a greater extent if they contradict each other rather than if one position includes a sentence for which the other remains silent. According to this choice, the Hamming distance is normalised accordingly by the doubled size of the sentence pool $2n$.

$1 - \Delta$ is a measure of similarity between two positions. The maximal value is 1, and it occurs if and only if the positions are identical. The minimal value of 0 comes about if and only if two complete positions contradict each other with respect to every sentence.

The two-point-ensemble is suitable to investigate compatibility and similarity, as the operationalised measures can be applied to the paired sets of initial commitments as well as the outputs. This leads to the following setup to extract results from the data: For pairs of sets of initial commitments we determine how many of them are dialectically compatible. Moreover, we calculate the similarity between positions of each pair of inputs.

As we have seen in the previous section, we need to account for the formal model producing multiple outputs per simulation setup. First, we form pairwise combinations between all outputs reached from the first and the second set of initial commitments from a pair of simulation setups.[6] After this pairing, we determine how many of the output pairs are compatible, and calculate the similarity between the position of each output pair.

So operationalised, the no-convergence objection would lead us to expect that agreement in terms of compatibility or similarity is not boosted or even reduced by RE.

Results Figure 11.3 and gives a visual impression of the following results concerning compatibility. Every subplot in this figure displays the relative shares of compatible input (left) and for global optima and fixed points for two configuration of weights, as well as the streamlining baseline. The relative share of compatible paired sets of initial commitments is rather low,

[6] The order does not matter as compatibility and similarity are symmetrical.

indicating that the randomly chosen sets of initial commitments are incompatible most of the time.[7]

The relative share of pairs of global optima and fixed points commitments is substantially boosted for the configuration in (A) and (B). Moreover, we can examine the "flow" between inputs and outputs. Most of the compatible pairs of initial commitments yield compatible pairs of outputs. Only a small portion of compatible input pairs yield incompatible pairs of outputs. Notably, a substantial amount of incompatible inputs yield compatible outputs.

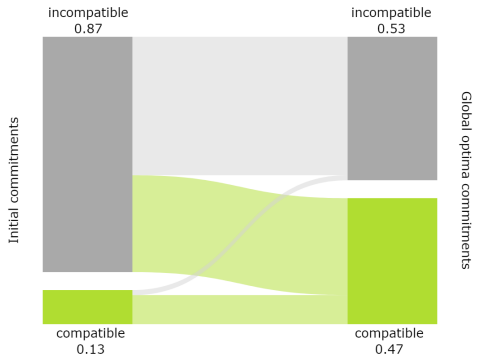
For the second configuration in (C) and (D), things look differently. The formal model is able to preserve the most part of compatible inputs, but it mostly fails to establish compatible outputs from incompatible inputs. Moreover, it falls short of the performance of the streamlining baseline.

For similarity, the outputs have been split into bins according to initial similarity, which allows to plot output similarity against these bins as depicted in Figure 11.4. The boxes cover the middle 50 percent of ordered values, the interquartile range (IQR). The whiskers attached to the box have a maximal length of $1.5 \times IQR$ (or are restricted to the most extreme actual values covered by them). Every value outside of the box and the whiskers is treated as an *outlier* represented by a dot. The horizontal line between the notches of a box indicates the median, the middle value in an ordered data set. It is more robust with respect to outliers and skew than the arithmetic mean. Notches indicate the 95% confidence interval, conveying a rough visual indicator for significant differences (McGill, Tukey, and Larsen, 1978).

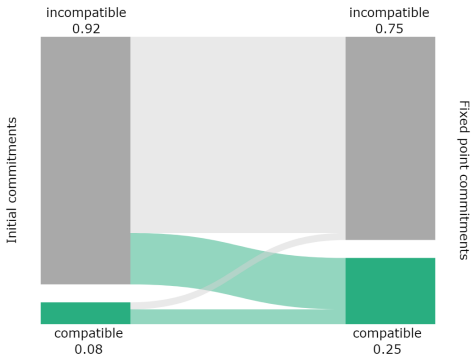
This leads to the following observations: The median similarity among pairs of global optima, fixed point or streamlined output commitments is above the dashed line, which indicates parity between initial and output similarity, for low and moderate values of initial similarity. This means that in these cases, output similarity is on average significantly higher than the similarity between the pairs of sets of initial commitments that served as inputs to produce them. Apart from the boost for low values of initial similarity, output similarity is roughly proportional to the initial similarity,

In order to provide a plot that allows for a better comparison between global optima, fixed points and streamlined outputs, I compose the results

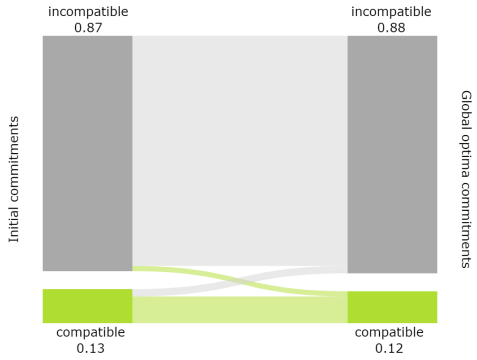
[7] Recall that the formal model does not require a set of initial commitments to be dialectically consistent. The relatively high share of incompatible pairs of sets of initial commitments is due to the fact that two positions are automatically incompatible if at least one of them is dialectically inconsistent. Differences in relative shares of compatible sets of initial commitments between are due to the fact that equilibration, global optimisation and streamlining may produce different numbers of outputs per simulation setup.



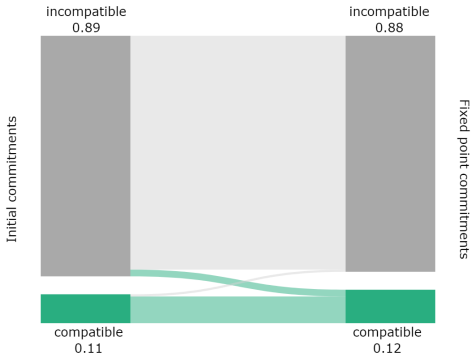
(A) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$



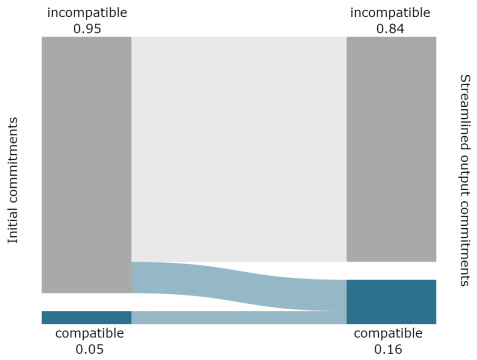
(B) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$



(C) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$



(D) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$



(E) Streamlining

FIGURE 11.3: Compatibility "flow" between pairs of inputs (left bar) and outputs (right bar) for global optima (light green), fixed points (turquoise) and the streamlined outputs (blue).

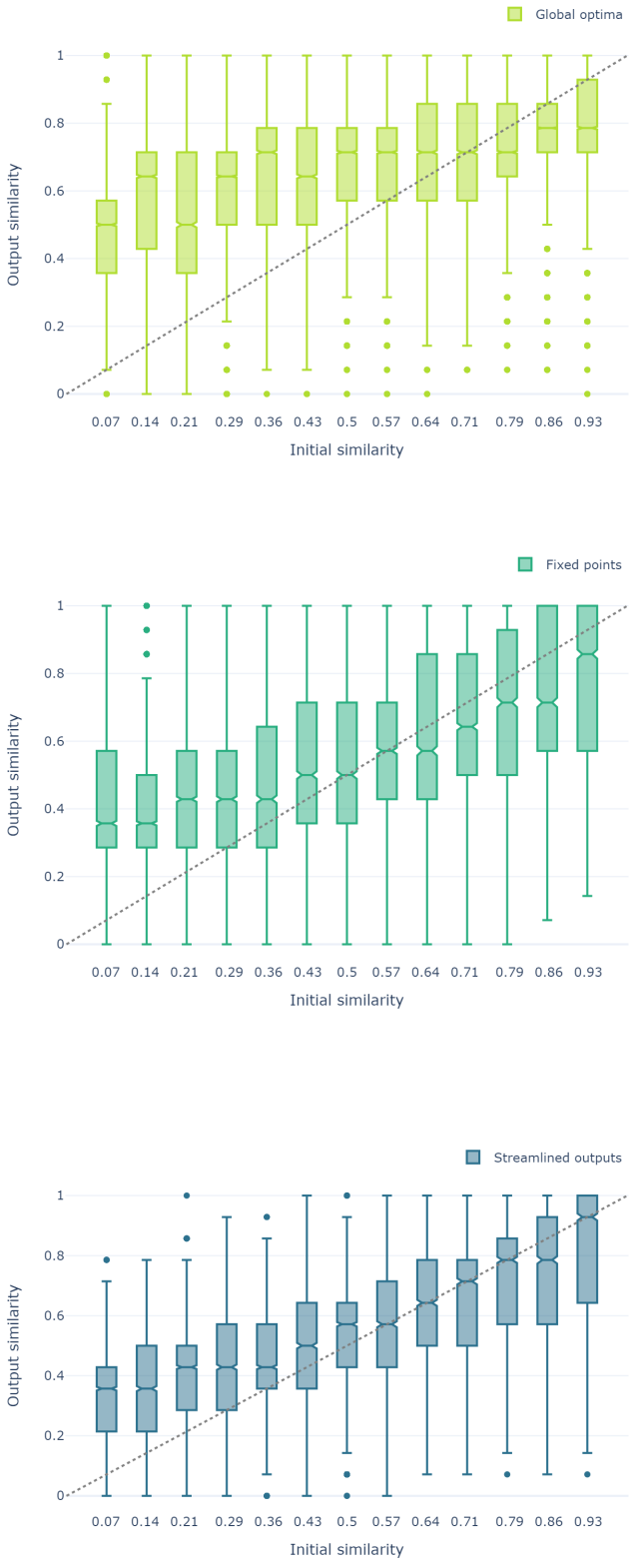


FIGURE 11.4: Plotting output similarity against initial similarity bins for global optima and fixed points for the configuration $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$, as well as for the streamlining baseline. The grey, dashed line indicates parity between input and output similarity.

differently in Figure 11.5, and complement them with the second paradigmatic configuration of weights. For the first configuration in (A), global optima perform better than the streamlining baseline except for very high values of initial similarity. Fixed points perform roughly equally to the baseline and fall short for high values of similarity. For the second configuration in (B), global optima and fixed points get outperformed by the streamlining baseline in many cases.

Discussion The formal model of RE promotes agreement to some extent. Concerning compatibility, the model is able to preserve compatibility from inputs, and to establish compatibility in a substantial amount of incompatible pairs of inputs for specific configurations of weights.

This result does not show that agents reach the same outputs. However, if their initial commitments are incompatible and they can aggregate their compatible sets of output commitments without running into contradictions, this can nonetheless be understood as a form of convergence. The agents reached agreement on the sentences that they both accept or reject. The remaining differences can be traced to commitments that one agent accepts or rejects, while the other agent remains silent on them.

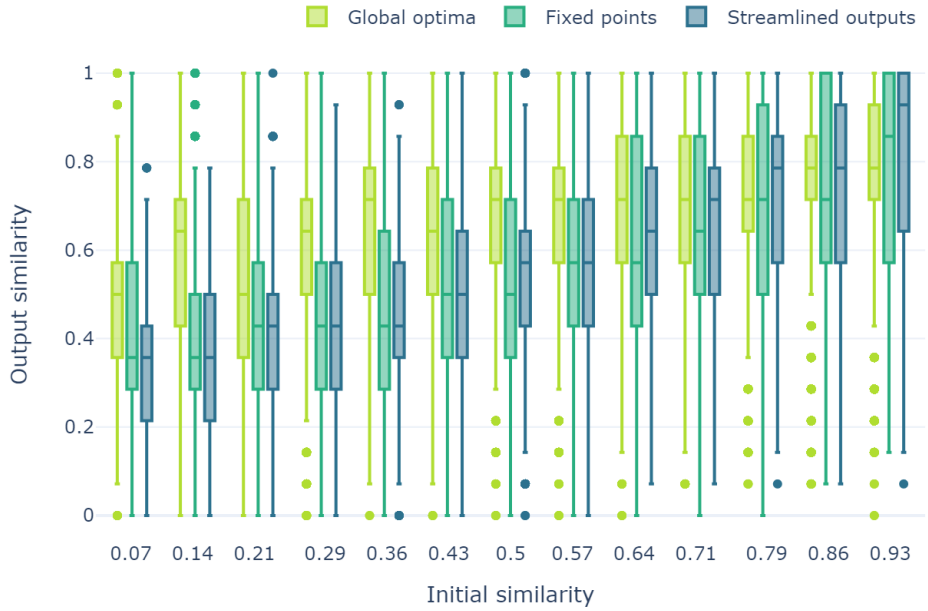
The small relative share of compatible pairs of inputs that lead to incompatible outputs arises from the following situation. There are cases in which the only theories that account for the union of the sets of initial commitments are highly unattractive according to the measure for systematicity. Consequently, both agents choose better-performing theories that are immediately incompatible with each other. The rest of the equilibration proceeds by adjustments of commitments that "pass down" the incompatibility to the commitments before the agents settle on their respective fixed points.

Consider the following pair of initial commitments: $C_0 = \{3, 4, 5\}$ and $C_0' = \{5, 6, 7\}$. C_0 and C_0' are compatible in the dialectical structure of the standard example. Equilibration processes (standard configuration of weights), which do not involve random choices, yield the following fixed points, which happen to be full RE states:

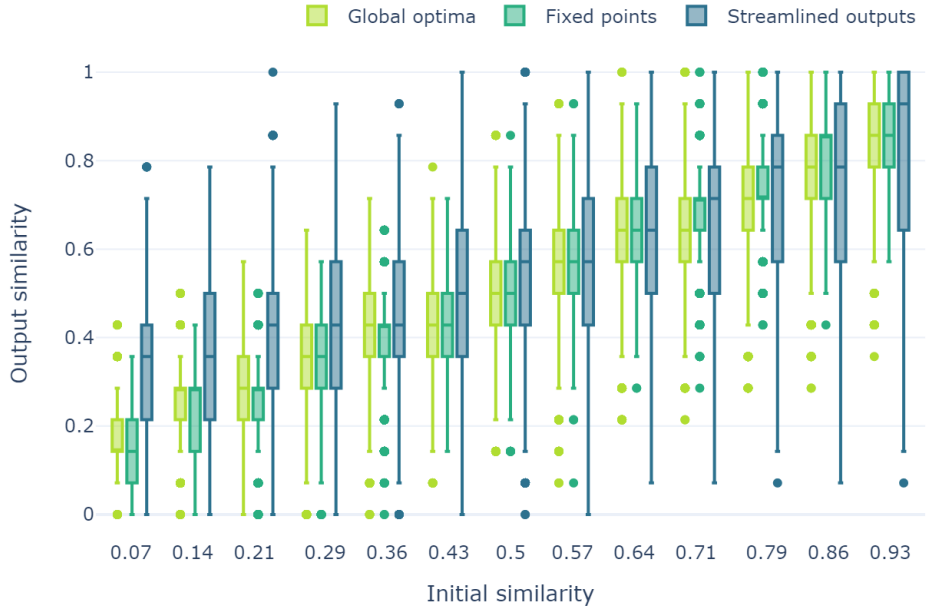
$$T = \{1\} \text{ and } C = \{1, 3, 4, 5, -6, -2\}$$

$$T' = \{2\} \text{ and } C' = \{2, 5, 6, 7, -4, -1\}$$

Despite compatible initial commitments, fixed point commitments are incompatible with each other (as well as the theories). Let us have a closer look



(A) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$



(B) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$

FIGURE 11.5: Plotting output similarity against initial similarity bins for global optima, fixed points and the streamlined outputs.

at the equilibration in an attempt to understand what is going on. According to the definition, C_0 and C_0' are compatible because there is a complete and consistent position that extends both In fact, they have a single complete extension in common, namely $\{3,4,5,6,7,-1,-2\}$. The only axiomatisation of this position that could serve as a serious theory candidate during the first adjustment step for both equilibration processes is $\{3,4,5,6,7\}$. However, this axiomatisation performs terribly with respect to systematicity. Consequently, both agents choose better performing theories that are immediately incompatible with each other. The rest of the equilibrations proceed by adjustments of commitments that "pass down" the incompatibility to the commitments before the agents settle on their respective fixed points.

This example illustrates that two compatible positions can agree on only one complete consistent extension. This weak tie may then be "severed" during an equilibration process, if said extension (or its axiomatisation) is unattractive according to the achievement function. Compatibility, as a categorical feature, does not portray the details of a setup. It requires the existence of a common, complete and consistent extension, but it does not capture to what degree the complete and consistent extensions of two positions overlap or the performance of positions in the overlap according to the achievement function. If such additional or even coincidental features of the setup become relevant only later, during the equilibration process or global optimisation, vanishing compatibility has to be expected to some extent.

Agreement, spelled out as similarity on the level of sentences, is on average slightly increased over inputs. The more agents start from similar initial commitments, the more they tend to reach similar outputs. This is more than what the no-convergence objections would lead us to expect. In comparison to Figure 3.2, if agents did not converge we would have expected that the results would tend to fall below the dashed line in Figure 11.4. Now, though, we must face the question whether this is sufficient agreement. This, however, will have to be a subject for future discussion in the informal debate about RE, as critics thus far are silent on this point.

I suppose that the inclusion of systematicity into the formal model explains why the formal model is able to boost agreement. Systematicity restricts candidate theories in adjustment steps to those which strike a good balance between containing few sentences (simplicity) and entailing many (scope). Account then serves to realise this potential in the output commitments, which have been studied. The restriction of candidate theories translates to reduced possibilities for incompatibility and dissimilarity. This is

underwritten by additional simulations for low values of systematicity and account in Appendix E.1.

This is an important additional result. Some proponents of RE take systematisation to be the "key driver" of equilibration (Baumberger and Brun, 2021, 7928), but it seems to me that critics often underestimate this aspect of RE.

11.4 Does Reflective Equilibrium Allow for "Anything Goes"?

Methods The full-spectrum-ensemble is suitable to study "anything goes" as it operationalises agents that start from maximally diverse initial commitments in a dialectical structure. I analyse whether "anything goes" holds in each of the 30 randomly generated structures separately, and subsequently report averages across the structures.

"Anything goes" on the level of sets can be operationalised straightforwardly as a comparison between the number of sets of initial commitments and the number of different sets of output commitments. If "anything goes" holds, we cannot expect to see a substantial reduction in numbers between inputs and outputs, as "there might be as many possibly conflicting, justified moral belief sets as there are people engaging in the method of reflective equilibrium" (de Maagt, 2017, 450).

How are we to check whether "anything goes" holds on the level of sentences? We start from an individual simulation setup and form the union of all commitments of outputs of a kind (global optima, fixed points, or streamlined outputs). In the union, we count the number of different pairs of contradicting sentences, i.e., the number of s_i ($i = 1, \ldots, n$), such that s_i and $\neg s_i$ are members of the union, where n is the size of the unnegated half of the sentence pool. If every sentence from the sentence pool as well as its negation occur at least once, we say that the outputs *cover* the entire sentence pool, in which case "anything goes" obtains on the level of sentences.

Take, for example, case (C) from (Beisbart, Betz, and Brun, 2021): $C_0 = \{3, 4, 5, 6, 7\}$ in the standard dialectical structure (see Figure 7.1 and Table 7.1) with the configuration $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$ yields

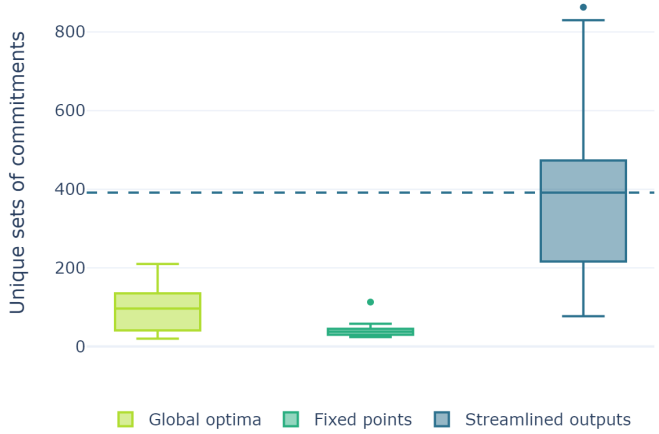
$$C\{1, 3, 4, 5, -6, -2\}, \{1\}) \text{ and } (\{2, 5, 6, 7, -4, -1\}, \{2\})$$

as global optima that perform equally well according to the achievement function. The union of their commitments is $\{1, -1, 2, -2, 3, 4, -4, 5, 6, -6, 7, \}$, and there are four pairs of contradicting sentences. Thus, even if the outputs do not cover the entire sentence pool, and hence do not exhibit "anything goes" on the level of sentences, there is substantial disagreement. We collect this information for all simulation setups in the full-spectrum-ensemble.

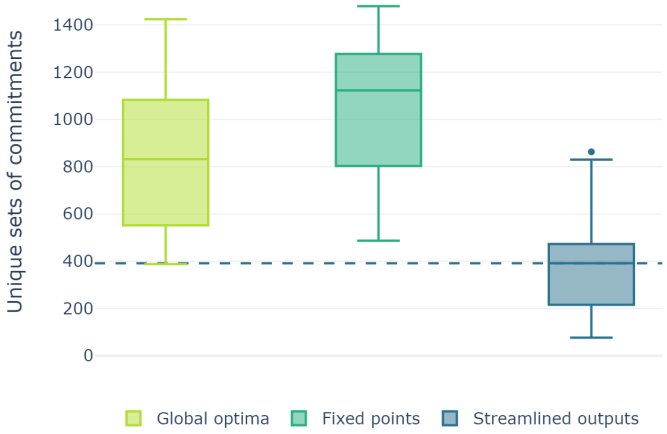
Of course, we could group simulation setups more inclusively, for example by taking the union over all outputs from the full spectrum. However, I take the present approach to be in line with literature on RE, especially Elgin's discussion of pluralism and relativism that ensues from RE (1996, 135–145). There, she discusses examples of incompatible systems that are "elaborations of the same initially tenable commitments" Elgin (1996, 136, 139). From this, I infer that it would be serious a problem if agents could frequently reach "anything goes" from the *same* set initial commitments, or more generally, from the *same* epistemic situation. Elgin (1996, 14, 142) doubts that this is the case. Initially tenable commitments, e.g., that exterminating a race is wrong, or that π is an irrational number, cannot be revised without compelling reasons. As it is not the case that "grounds for radical revision are generally available", this puts severe constraints on what can result as a tenable system from the same set of initially tenable commitments.

Results On the level of sets, a drastic reduction between the number of input and output positions is apparent. Figure 11.6 displays the number of different sets of output commitments reached from the full spectrum of 2186 sets of initial commitments in 30 randomly generated structures. For the first configuration in (A), the median number of different sets of global optima commitments is 96 (IQR: 41–134), and for fixed points the median number is 38 (IQR: 30–45). This is significantly better than the streamlining procedure with a median number of different sets of output commitments of 392 (IQR: 222–471). Note that the vertical axis is readjusted for the second configuration (B). Global optima (median: 832 , IQR: 578–1082) and fixed points (median: 1122, IQR: 842–1276) reach significantly more different sets of output commitments than the streamlining baseline.

Let us turn to "anything goes" on the level of sentences. Figure 11.7, Figure 11.8 and Figure 11.9 present the results for global optima, fixed points and streamlined outputs, respectively. The 65,580 simulation setups per configuration have been grouped horizontally according to the size of the union of their output commitments. The count of simulation setups is depicted



(A) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$



(B) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$

FIGURE 11.6: Number of different sets of output commitments reached from the full spectrum of 2186 sets of initial commitments in 30 randomly generated structures.

vertically and coloured according to the number of contradicting pairs of sentences in the union of output commitments of a simulation setup.

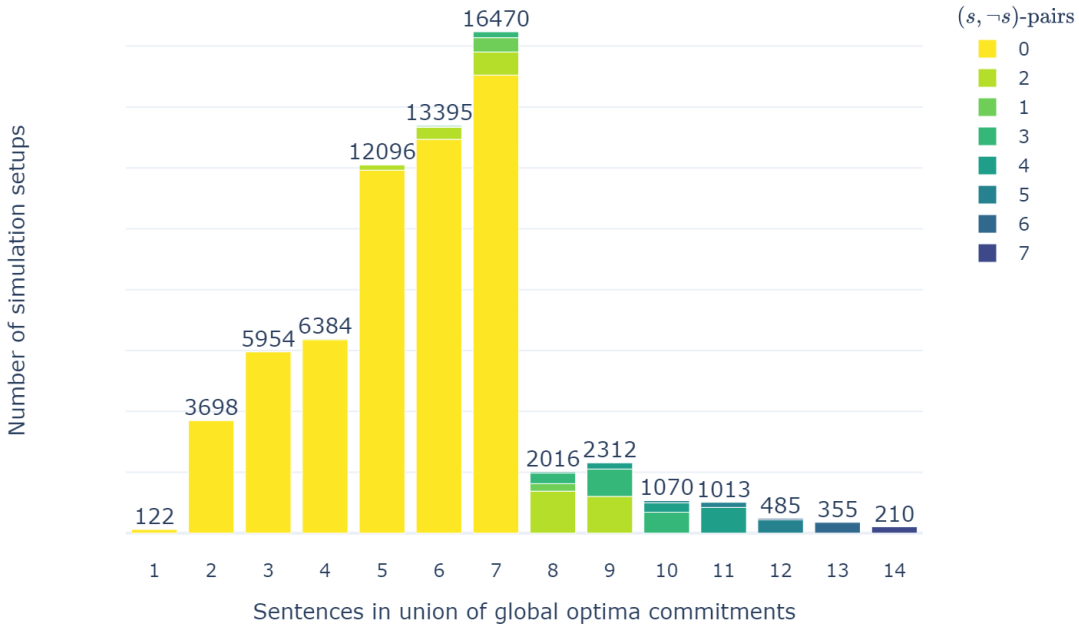
In all figures, the vast majority of simulation setups result in a union of output commitments that do not include a single pair of contradicting sentences (yellow). Note that the full-spectrum-ensemble consists of dialectical structures with seven unnegated sentences. Thus, for eight or more sentences in the union of output commitments from a simulation setup, there is at least one pair of contradicting sentences. At this point, there is a significant drop in the number of simulation setups.

"Anything goes", in the most strict sense, applies if there are simulation setups for which the outputs cover all 14 sentences of the sentence pool (rightmost bar). These cases occur very rarely. For example, there are 210 (0.3% of 65,580) simulation setups that result in global optima commitments covering the entire sentence pool for configuration (A) in Figure 11.7.

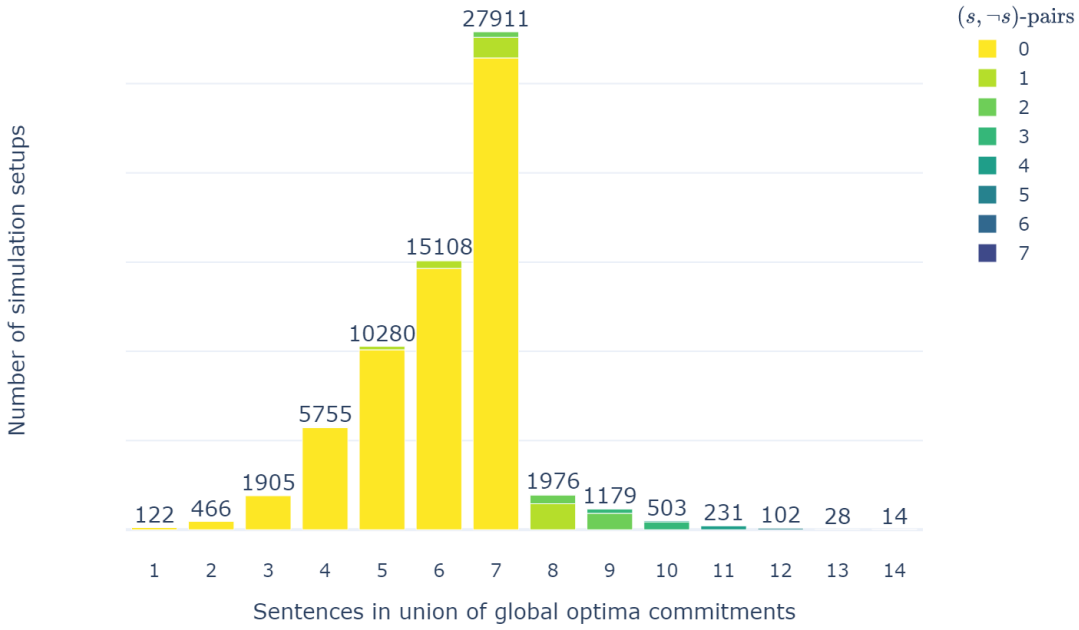
For both configurations, global optima and fixed points perform mostly better than the streamlining baseline with respect to contradicting sentences among the outputs reached from a simulation setup, i.e., the right half of the plots, especially for 8–12 sentences in the union of output commitments. In general, there are fewer simulation setups that result in contradicting outputs.

As an exception to many other simulation study results, the model performs better for the configuration in (B) than for the configuration in (A) in Figure 11.7 and Figure 11.8. I suppose that this can be traced to the conservativity exhibited by the model for configuration of weights that put more weight on faithfulness than on account (see Chapter 10). Thus, the preservation of sets of initial commitments, which are required to be minimally consistent, and hence contain no contradicting pairs of sentences, may explain the good performance.

Discussion The formal model drastically reduces the number of sets of commitments in comparison to the number of sets of initial commitments. Requirements and desiderata that are drawn from informal accounts of RE and implemented in the formal model thus sift out positions drastically. At least, if they are configured appropriately. Giving too much weight to faithfulness renders the model conservative (Chapter 10), and here we can observe that this leads to substantially more different sets of output commitments from the full spectrum of sets of initial commitments. If, in contrast, the theoretical virtues of account and systematicity receive more weight, the

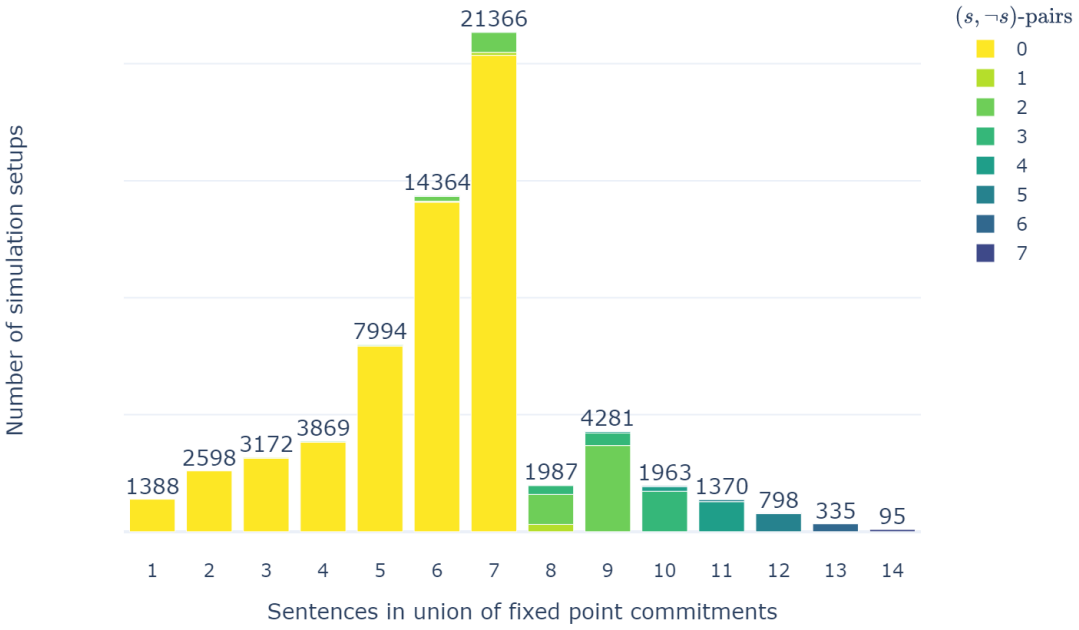


(A) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$

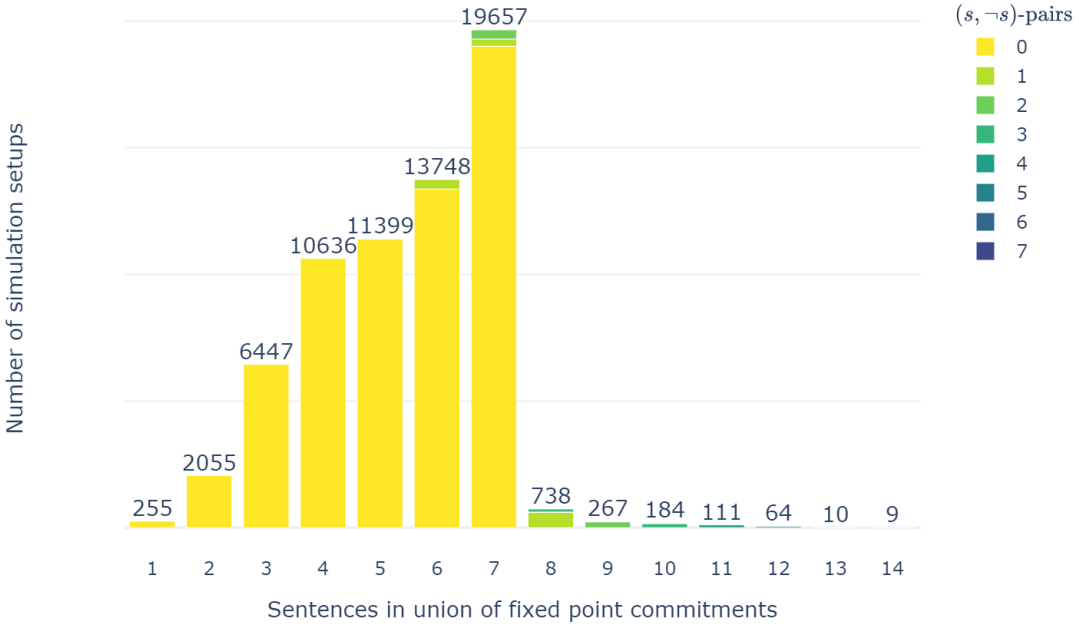


(B) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$

FIGURE 11.7: Number of simulation setups (vertical axis) grouped by the number of different sentences in the union of global optimum commitments (horizontal axis). Colours indicate the number of different pairs of contradicting sentences in the union of commitments.



(A) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$



(B) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$

FIGURE 11.8: Number of simulation setups (vertical axis) grouped by the number of different sentences in the union of fixed point commitments (horizontal axis). Colours indicate the number of different pairs of contradicting sentences in the union of commitments.

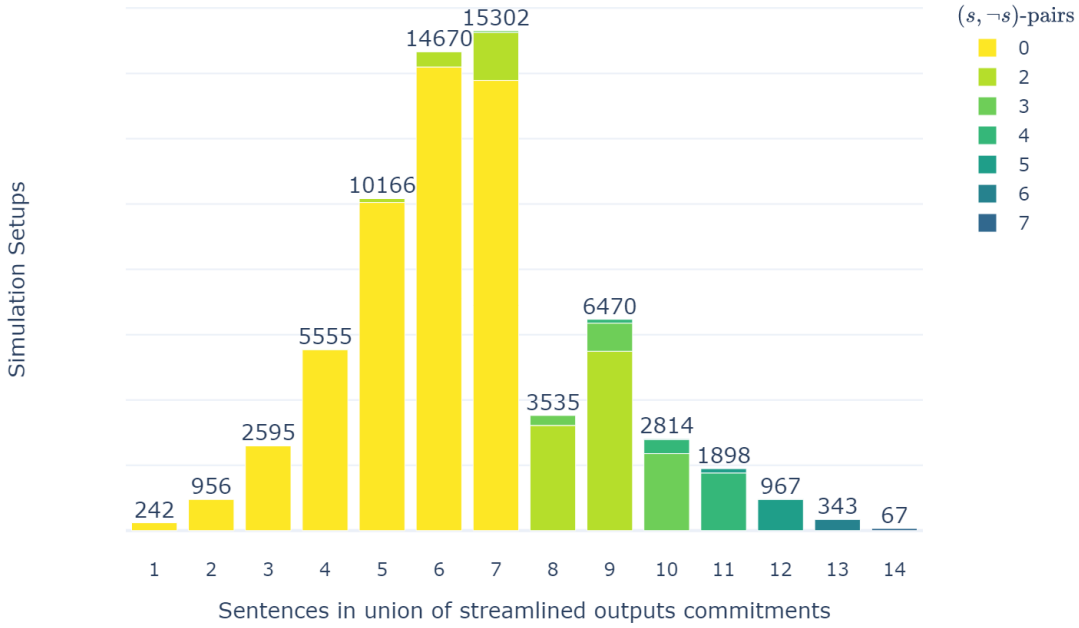


FIGURE 11.9: Number of simulation setups (vertical axis) grouped by the number of different sentences in the union of streamlined output commitments (horizontal axis). Colours indicate the number of different pairs of contradicting sentences in the union of commitments.

model reduces over two thousand different sets of initial commitments to a few dozen sets of output commitments .

I suppose that, the existence of a configuration of weights that leads the model of RE to perform that well is enough to dispel DeMaagt's worry that there might be as many outputs as there are inputs. "Anything goes" can be evaded on the level of sets if theoretical virtues are given due consideration in RE.

On the level of sentences, "anything goes", in the strictest sense of covering the entire sentence pool, obtains very rarely in ten thousands of simulation setups. These rare cases of "anything goes" from an individual simulation setup in the full spectrum-ensemble would require a detailed analysis to see whether they are a mere artefact of the random generation of dialectical structures or extremely "unfortunate" sets of initial commitments.

As it stands, the model does not track the tenability (Elgin, 1996), or the independent credibility (Baumberger and Brun, 2021) of (initial) commitments, which might further reduce the range of admissible inputs, outputs or adjustments. Additional research in this direction may prove to be insightful, but it goes beyond the scope of the present project.

Still, the rare of occurrence of "anything goes" may also be due to the highly fine-grained grouping of outputs from individual simulation setups. This amounts to the presupposition of full agreement at the outset of RE. If we formed unions differently, for example, from all outputs from the full spectrum instead of individual sets of initial commitments, the unions would cover the entire sentence pool in every dialectical structure from the full-spectrum-ensemble. So, for sufficiently diverse sets of initial commitments the model produce outputs that cover the entire sentence pool, and hence allow for "anything goes" on the level of sentences.

Let me explain why I think that such results are not that problematic for RE. In case of "anything goes" from sufficiently diverse inputs, differences in equilibria can be traced to differences in the epistemic situations of agents, in particular to different sets of initial commitments. In contrast to the equilibration process in the formal model, which terminates whenever the stopping condition is met, even "wider" RE does not stop there. Scanlon (2003, 152f) and Tersman (2018, 7) stress the importance of taking known disagreements among different agents into account as they may disrupt the ever-provisional equilibria. If a group of agents reaches drastically different outputs, they should be suspicious of whether they all are in a state of equilibrium, and, hence, evaluate their current state in view of the others. This may lead to

further revisions.

The same line of thought applies to the other elements of simulation setups that have been presented in a way that presupposes agreement among agents with respect to dialectical structures and configuration of weights. It is quite plausible that the model would produce more diverse outputs if we collected outputs across different dialectical structures in the background or configuration of weights. However, differences in outputs that can be traced to differences in simulation setups/epistemic situations (or different adjustment decisions during equilibration) present new input.

I take the following to be a lesson that we can draw from this: Even though that the instructions of a method of RE may be applicable by an individual epistemic agent, it is not recommendable to work in complete isolation from other agents. It is this kind of isolation which would free them of having to react to other views. Elgin (1996, 111–119) and Rechnitzer (2022, 34f) highlight that agents need to "go public" by relying on or taking into consideration the background or the commitments of others.

The present model is individualistic for the sake of simplicity. It does not allow agents to interact with each other during equilibration or to react to reaching different outputs. This opens up a series of interesting questions for further research in formal models of multi-agent RE. Which mechanisms can model such interactions or reactions? Do they lead to more consensus or polarisation among groups of agents?

11.5 Conclusion

Exploring simulations has revealed that the formal model of RE does not behave as no-convergence objections would lead us to expect. Moreover, the results meet the expectations of proponents of RE:

> [Wide reflective equilibrium] has resources that might lead inquirers toward a greater degree of agreement. Nevertheless, it seems most reasonable to expect that, in the end, [wide reflective equilibrium] will produce convergence upon a small number of alternative moral views with significant differences rather than convergence on a single view.(DePaul, 2013, 4474)

The formal model does not always reach intra- or interpersonal convergence to a unique output, but it promotes agreement to some extent, and the threat of "anything goes" can be kept at bay effectively.

I take this to be good news for proponents of RE. For the first time in the debate about convergence in RE, we can go beyond speculation and back up plausibility considerations by reference to computer-generated data. I cannot see a reason to think that these results are a mere artefact of formalisation, and they stem from a formal model that carefully takes up components of elaborate informal accounts of RE.

I draw the following lessons for the informal debate about RE and the possibility of providing an elaborate account of RE that is defensible against no-convergence objections in view of the present results. First, the failure to produce a unique output motivates us to adopt a pluralist stance on justification with RE. Next, the inclusion of systematicity, i.e., the consideration of theoretical virtues in RE, proved to be convergence-conducive. The demand for systematisation is more or less implicit in classic and elaborate accounts of RE, but it seems that it often escapes critics' notice. While "systematic" or cognate terms are mentioned explicitly, these notions are often not further spelled out, and they do not play a tangible role in equilibration. Finally, simulation setups reflect the epistemic situation of an agent that engages in RE. It is a merit of RE that it forces us to be explicit about such things (Rechnitzer, 2022, 241), and I propose to report equilibria relative to epistemic situations. Simulations indicate that "anything goes" arises rarely from single setups, and more frequently from collections of diverse setups. Taking differences in epistemic situations into account, however, may provoke further revisions, which keeps the threat of "anything goes" at bay.

Naturally, the present study faces limitations. It rests on examples with a very small sentence pool, for example. Unfortunately, the search space grows exponentially in the number of sentences, and thus computational feasibility is quickly exhausted. This may be mitigated by switching from the costly semi-global optimisation of the present model to locally searching variants. Such variants can handle much larger number of sentences, and they model agents that proceed in a "piecemeal" fashion, another under-explored idea that originates from Goodman (1983(1955)). However, larger sentence pool sizes prevent us from achieving global optimisation, and hence we must forfeit the determination of full RE states.

Next, the formal model does not pull out all the stops. As it stands, the model does not track the tenability of initial commitments (Elgin, 1996), the independent credibility of commitments (Baumberger and Brun, 2021), or interactions between agents (Tersman, 2018), which can further reduce the range of admissible inputs, outputs or adjustments, respectively. All of this

is clearly the object of further research.

Appendix

E.1 Robustness

E.1.1 Configuration of Weights

The two-point ensemble has been generated with seven weight configurations, of which two have been presented in the chapter. The following tables cover the results from all configurations. Note that the streamlining procedure does not involve a configuration of weights, and hence always yields the same results.

configuration	global optima	fixed points	streamlining
(0.35, 0.55, 0.10)	0.33	0.61	0.64
(0.55, 0.35, 0.10)	0.61	0.65	0.64
(0.70, 0.20, 0.10)	0.65	0.65	0.64
(0.55, 0.20, 0.25)	0.44	0.55	0.64
(0.46, 0.10, 0.44)	0.59	0.64	0.64
(0.10, 0.55, 0.35)	0.33	0.41	0.64
(0.10, 0.35, 0.55)	0.61	0.68	0.64

TABLE E.1: Intrapersonal convergence: relative share of simulation setups that result in a unique output grouped by configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$. The number of simulation setups per configuration is 26,000.

Table E.1 covers intrapersonal convergence operationalised as yielding a unique output from a simulation setup. In many cases, the outputs of the formal model perform about equally well as the streamlining baseline. However, there is a notably lower relative share of simulation setups with a unique global optimum for $\alpha_S = 0.55$. I suspect that this is again due to the proliferation of trivial output states (singleton theory and a single commitment) if systematicity gets a lot of weight.

Concerning interpersonal convergence from two sets of initial commitments to a unique output, Table E.2 presents the absolute number of pairs of simulation setups that both yield a unique output, as well as the relative

	global optima		fixed points		streamlining	
configuration	count	share	count	share	count	share
(0.35, 0.55, 0.10)	1803	0.39	5189	0.19	5621	0.05
(0.55, 0.35, 0.10)	5248	0.27	5800	0.19	5621	0.05
(0.70, 0.20, 0.10)	5915	0.25	5951	0.14	5621	0.05
(0.55, 0.20, 0.25)	2899	0.25	4376	0.15	5621	0.05
(0.46, 0.10, 0.44)	4829	0.16	5648	0.11	5621	0.05
(0.10, 0.55, 0.35)	1803	0.07	2686	0.05	5621	0.05
(0.10, 0.35, 0.55)	5240	0.05	6367	0.03	5621	0.05

TABLE E.2: Interpersonal convergence: Number of pairs of simulation setups that both yield a unique output, and the relative share of those which both reach the same unique output grouped by configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$. The total number of pairs of simulation setups per configuration is 13,000.

share among those that yield the same unique output. Note that the total number of simulation setups per configurations is 13,000. There is a significantly reduced number of simulation setups that both yield a unique global optimum for configurations with $\alpha_S = 0.55$. Here, too, the proliferation of trivial states for high values of α_S probably results in multiple outputs in many cases. Configurations of weights that put more weight on faithfulness than on account do not let the model perform substantially better than the streamlining baseline.

	global optima		fixed points		streamlining	
configuration	input	output	input	output	input	output
(0.35, 0.55, 0.10)	0.10	0.42	0.09	0.30	0.05	0.16
(0.55, 0.35, 0.10)	0.13	0.47	0.08	0.25	0.05	0.16
(0.70, 0.20, 0.10)	0.13	0.37	0.08	0.20	0.05	0.16
(0.55, 0.20, 0.25)	0.08	0.11	0.08	0.13	0.05	0.16
(0.46, 0.10, 0.44)	0.06	0.08	0.07	0.11	0.05	0.16
(0.10, 0.55, 0.35)	0.10	0.13	0.09	0.14	0.05	0.16
(0.10, 0.35, 0.55)	0.13	0.12	0.11	0.12	0.05	0.16

TABLE E.3: Relative share of compatible pairs of sets of initial commitments and compatible pairs of sets of output commitments grouped by configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$.

Table E.3 collects results concerning the compatibility of paired sets of initial commitments as well as paired sets of output commitments. Note that the different relative shares of compatible inputs results from global optimisation, equilibration and streamlining producing different numbers of outputs per simulation setup. Overall, almost all outputs achieve to improve upon the relative share of compatible pairs but to very different extents. If faithfulness receives more weight than systematicity or account, the boost in compatibility is rather small. For (0.10, 0.35, 0.55), there is even a reduction in compatibility for global optima.

Table E.4 and E.5 display the results for similarity for global optima and fixed points respectively. Without going too much into the details, I suppose that the general picture presented in the chapter holds. Output similarity depends on the initial similarity. Global optima and fixed points improve upon low initial similarity, but result in a reduction if the initial similarity is high. Global optima tend to perform better than fixed points, and there are configuration of weights, for which the model performs better than the streamlining baseline for the most part.

	initial similarity												
configuration	0.07	0.14	0.21	0.29	0.36	0.43	0.50	0.57	0.64	0.71	0.79	0.86	0.93
(0.10, 0.35, 0.55)	0.14	0.29	0.29	0.36	0.43	0.43	0.50	0.57	0.64	0.64	0.71	0.79	0.86
(0.10, 0.55, 0.35)	0.21	0.29	0.36	0.43	0.50	0.50	0.57	0.57	0.64	0.71	0.71	0.79	0.79
(0.35, 0.55, 0.10)	0.57	0.64	0.57	0.64	0.64	0.64	0.64	0.71	0.71	0.71	0.71	0.79	0.79
(0.46, 0.10, 0.44)	0.29	0.29	0.43	0.43	0.43	0.43	0.50	0.57	0.57	0.57	0.71	0.71	0.79
(0.55, 0.20, 0.25)	0.29	0.29	0.36	0.43	0.43	0.43	0.50	0.57	0.57	0.57	0.71	0.71	0.79
(0.55, 0.35, 0.10)	0.50	0.64	0.50	0.64	0.71	0.64	0.71	0.71	0.71	0.71	0.71	0.79	0.79
(0.70, 0.20, 0.10)	0.36	0.43	0.43	0.43	0.50	0.50	0.57	0.57	0.64	0.71	0.71	0.79	0.86
streamlining	0.36	0.36	0.43	0.43	0.43	0.5	0.57	0.57	0.64	0.71	0.79	0.79	0.93

TABLE E.4: Median similarity between pairs of global optima sets of commitments grouped by similarity of corresponding pairs of sets of initial commitments (columns) and configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$ (rows).

Table E.7 and E.8 collect information about "anything goes" on the level of sentences for global optima and fixed points, respectively. There are 65,580 simulation setups per configuration. Note that for 8 or more sentences, there is at least one pair of contradicting sentences. At this point, there is a significant drop in the number of simulation setups that yield outputs with at least one pair of contradicting sentences.

"Anything goes", in the most strict sense, applies if there are simulation setups for which the outputs cover all 14 sentences of the sentence pool

	initial similarity												
configuration	0.07	0.14	0.21	0.29	0.36	0.43	0.50	0.57	0.64	0.71	0.79	0.86	0.93
(0.10, 0.35, 0.55)	0.14	0.29	0.29	0.36	0.43	0.43	0.50	0.57	0.64	0.71	0.71	0.86	0.86
(0.10, 0.55, 0.35)	0.21	0.29	0.36	0.43	0.50	0.57	0.57	0.57	0.64	0.71	0.79	0.79	0.79
(0.35, 0.55, 0.10)	0.43	0.43	0.43	0.50	0.57	0.57	0.57	0.64	0.64	0.71	0.71	0.71	0.79
(0.46, 0.10, 0.44)	0.29	0.29	0.36	0.36	0.43	0.43	0.50	0.57	0.57	0.64	0.71	0.79	0.86
(0.55, 0.20, 0.25)	0.29	0.29	0.36	0.36	0.43	0.43	0.50	0.57	0.57	0.64	0.71	0.79	0.86
(0.55, 0.35, 0.10)	0.36	0.36	0.43	0.43	0.43	0.50	0.50	0.57	0.57	0.64	0.71	0.71	0.86
(0.70, 0.20, 0.10)	0.36	0.36	0.43	0.43	0.43	0.43	0.50	0.57	0.57	0.64	0.71	0.79	0.86
streamlining	0.36	0.36	0.43	0.43	0.43	0.5	0.57	0.57	0.64	0.71	0.79	0.79	0.93

TABLE E.5: Median similarity between pairs of fixed point sets of commitments grouped by similarity of corresponding pairs of sets of initial commitments (columns) and configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$ (rows).

	global optima		fixed points		streamlining	
configuration	median	IQR	median	IQR	median	IQR
(0.10, 0.35, 0.55)	832	578 – 1082	1122	842 – 1276	392	222 – 471
(0.10, 0.55, 0.35)	1116	742 – 1422	1074	747 – 1384	392	222 – 471
(0.35, 0.55, 0.10)	156	130 – 168	174	106 – 191	392	222 – 471
(0.46, 0.10, 0.44)	156	126 – 177	206	194 – 226	392	222 – 471
(0.55, 0.20, 0.25)	175	128 – 201	200	186 – 209	392	222 – 471
(0.55, 0.35, 0.10)	96	41 – 134	38	30 – 45	392	222 – 471
(0.70, 0.20, 0.10)	32	28 – 36	43	38 – 48	392	222 – 471

TABLE E.6: Number of different sets of output commitments reached from the full spectrum of sets of initial commitments (2186 positions) in 30 randomly generated structures grouped by configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$.

	Number of sentences in union of global optima commitments													
configuration	1	2	3	4	5	6	7	8	9	10	11	12	13	14
(0.10, 0.35, 0.55)	122	466	1905	5755	10280	15108	27911	1976	1179	503	231	102	28	14
(0.10, 0.55, 0.35)	122	466	1910	5680	10598	14240	18077	5479	3853	2710	1365	709	313	58
(0.35, 0.55, 0.10)	122	2612	4766	5450	9874	12617	14422	5353	3945	3037	1838	985	412	147
(0.46, 0.10, 0.44)	122	304	385	1944	4425	11030	30762	2098	7528	3449	1816	1068	518	131
(0.55, 0.20, 0.25)	122	914	1671	2675	7354	9234	22900	2828	9698	4137	2016	1137	664	230
(0.55, 0.35, 0.10)	122	3698	5954	6384	12096	13395	16470	2016	2312	1070	1013	485	355	210
(0.70, 0.20, 0.10)	122	2373	3589	4484	10675	15184	20691	1813	3356	1102	1065	509	397	220
streamlining	242	956	2595	5555	10166	14670	15302	3535	6470	2814	1898	967	343	67

TABLE E.7: Number of simulation setups that yield a specific number of sentences across all global optima commitments reached from an individual simulation setup.

	Number of sentences in union of fixed point commitments													
configuration	1	2	3	4	5	6	7	8	9	10	11	12	13	14
(0.10, 0.35, 0.55)	255	2055	6447	10636	11399	13748	19657	738	267	184	111	64	10	9
(0.10, 0.55, 0.35)	255	2055	6447	9506	12387	13535	13884	3669	1713	1532	416	101	54	26
(0.35, 0.55, 0.10)	1970	5701	5321	4925	11977	11471	15474	1611	3080	1740	929	763	457	161
(0.46, 0.10, 0.44)	1170	1302	668	2362	5268	13725	29214	2841	5567	2203	899	279	68	14
(0.55, 0.20, 0.25)	1170	1906	1013	1983	7135	11516	27964	2818	5775	2457	1175	494	146	28
(0.55, 0.35, 0.10)	1388	2598	3172	3869	7994	14364	21366	1987	4281	1963	1370	798	335	95
(0.70, 0.20, 0.10)	1171	1602	686	2776	5709	14655	28491	2141	4645	1742	1171	527	208	56
streamlining	242	956	2595	5555	10166	14670	15302	3535	6470	2814	1898	967	343	67

TABLE E.8: Number of simulation setups that yield a specific number of sentences across all fixed points commitments reached from an individual simulation setup.

(rightmost column). These cases occur very rarely, and they are slightly more pronounced for global optima than for fixed points.

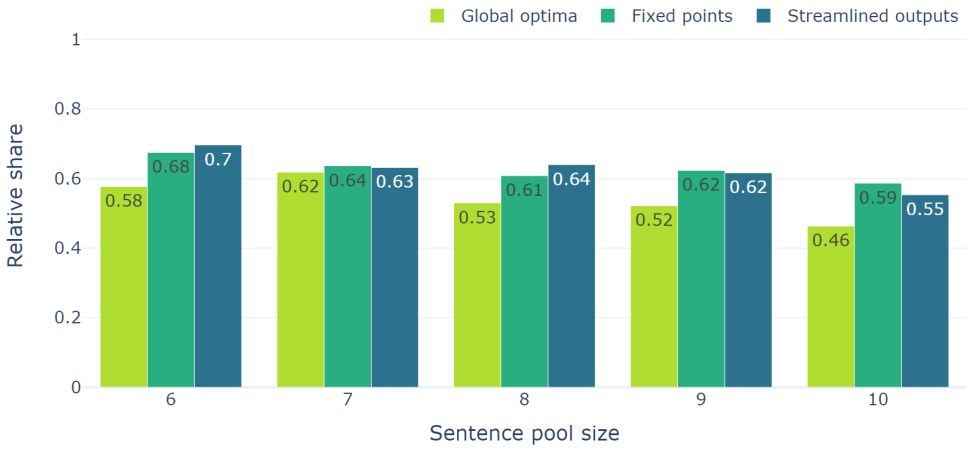
There are configuration of weights for which the model performs mostly better than the streamlining baseline, with respect to contradicting sentences among the outputs reached from a simulation setup (i.e., right half of the tables).

Interestingly, the model performs quite well for configuration of weights that put more weight on faithfulness than on account. We have observed earlier that these configuration lead to conservative behaviour (see Chapter 10). Thus, the preservation of sets of initial commitments, which are required to be minimally consistent (no contradicting pairs of sentences), may explain the good performance.

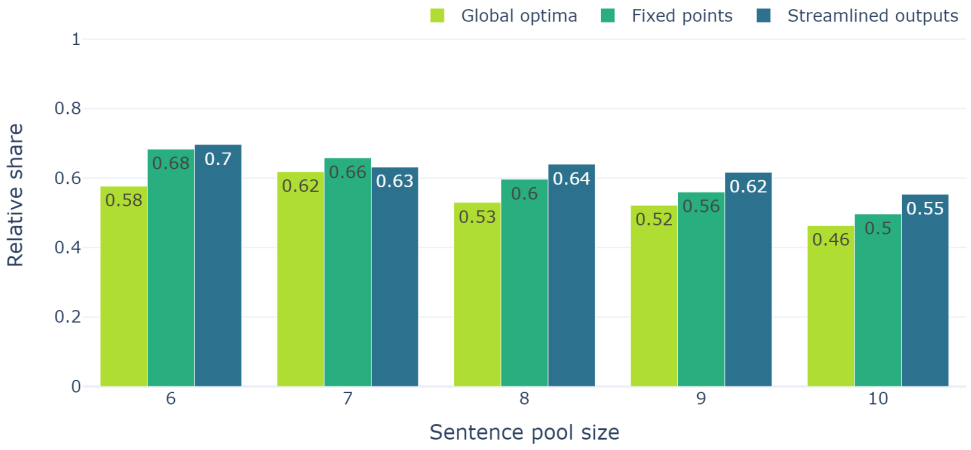
E.1.2 Sentence Pool Size

An additional ensemble serves to investigate the influence of the size of the sentence pool with respect to some aspects of convergence. Unfortunately, this does not include full-spectrum data due to computational limitations. Thus, I have to exclude "anything goes" from the following analysis. The ensemble includes two paradigmatic configurations of weights, sentence pools ranging from 6 to 10 sentences, and 300 randomly generated dialectical structures per sentence pool size. Per dialectical structure, a pair of random sets of initial commitments has been chosen such that the similarity of sets of initial commitments distributes evenly across low, medium or high values. This results in 6,000 simulation setups.

Figure E.1 and E.2 present the results for intrapersonal, and respectively, interpersonal convergence to a unique output. Concerning intrapersonal

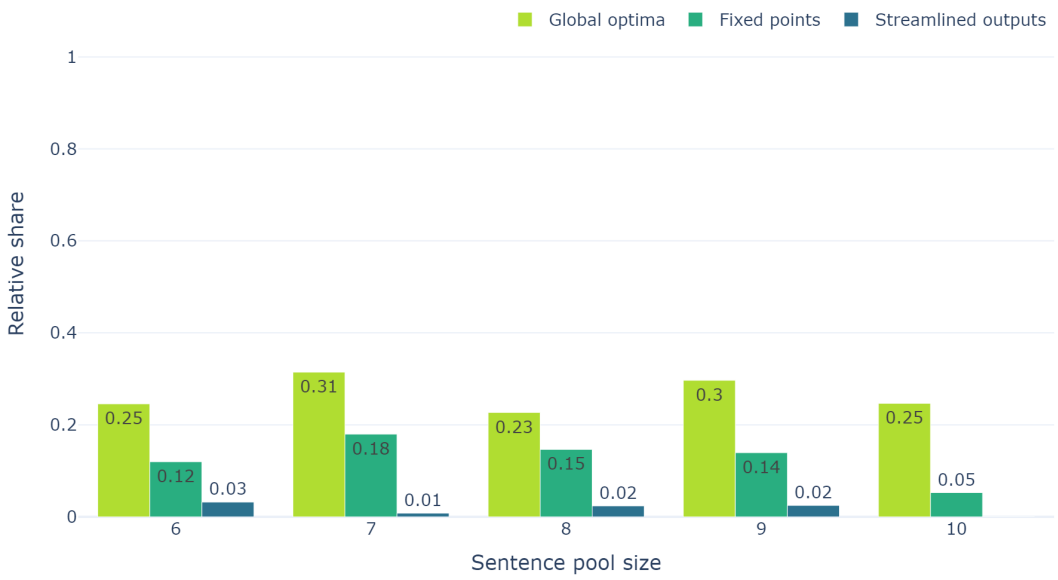


(A) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$

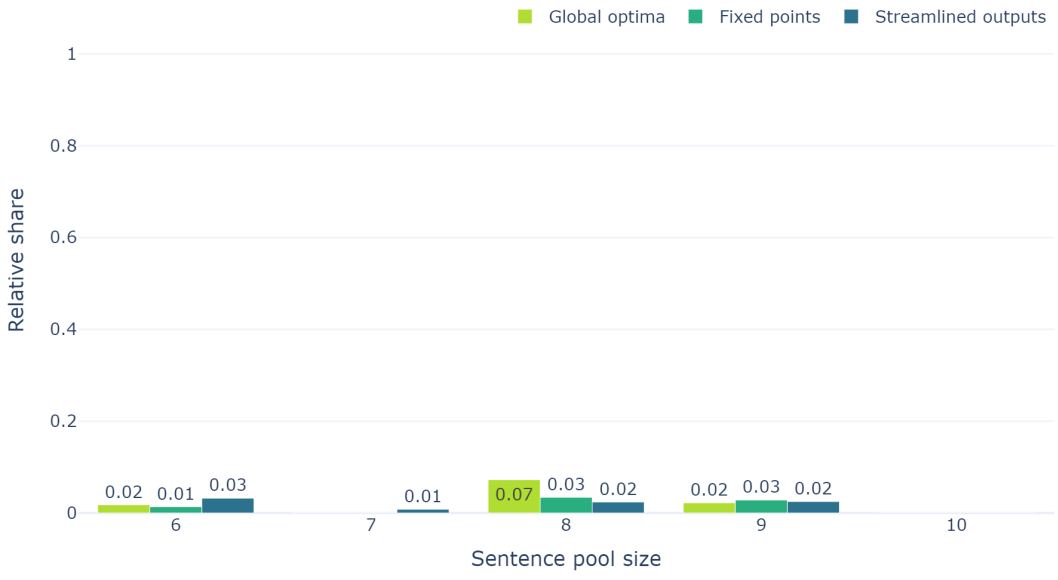


(B) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$

FIGURE E.1: Relative share of simulation setups that result in a unique output grouped by sentence pool. The total number of simulation setups per configuration and sentence pool size is 600.



(A) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$



(B) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$

FIGURE E.2: Relative share of paired simulation that result in the same, unique output grouped by sentence pool size.

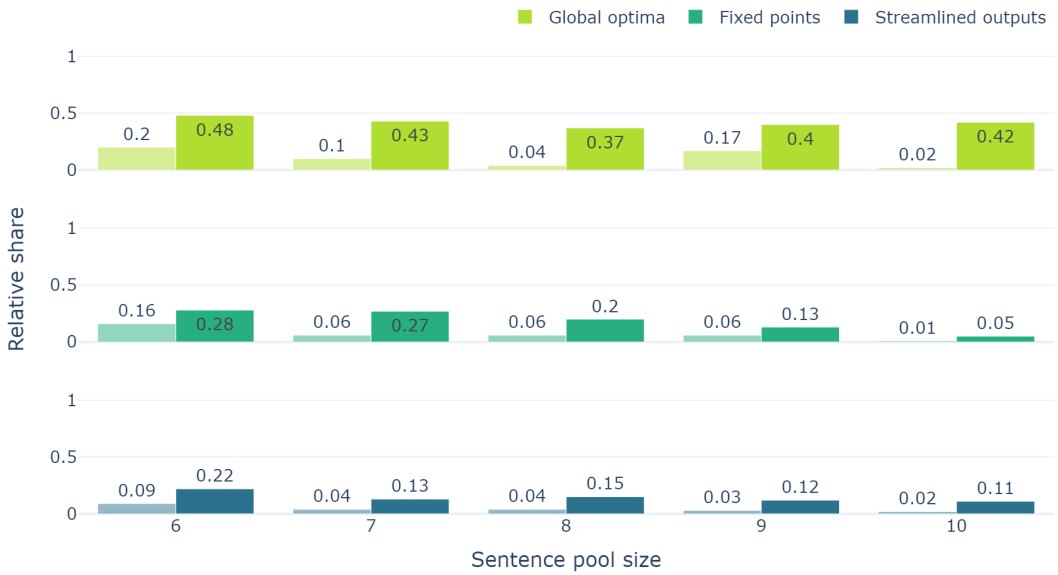
convergence in Figure E.1, fixed points and streamlined outputs perform roughly equally well and slightly better than global optima. In view of increasing sentence pool sizes, there is a weak decrease in the relative share of simulation setups that yield a unique output. Given the exponential growth (with base 3) of positions with increasing numbers of sentences, this is to be expected. There is no notable difference between the configuration of weights.

For interpersonal convergence in Figure E.2, global optima perform better than fixed points, which in turn, outrun streamlined outputs. With respect to the sentence pool size no clear trends are notable. The results for the second configuration of weights (B) cannot be interpreted meaningfully apart from the drastic reduction of the relative share in comparison to the first configuration.

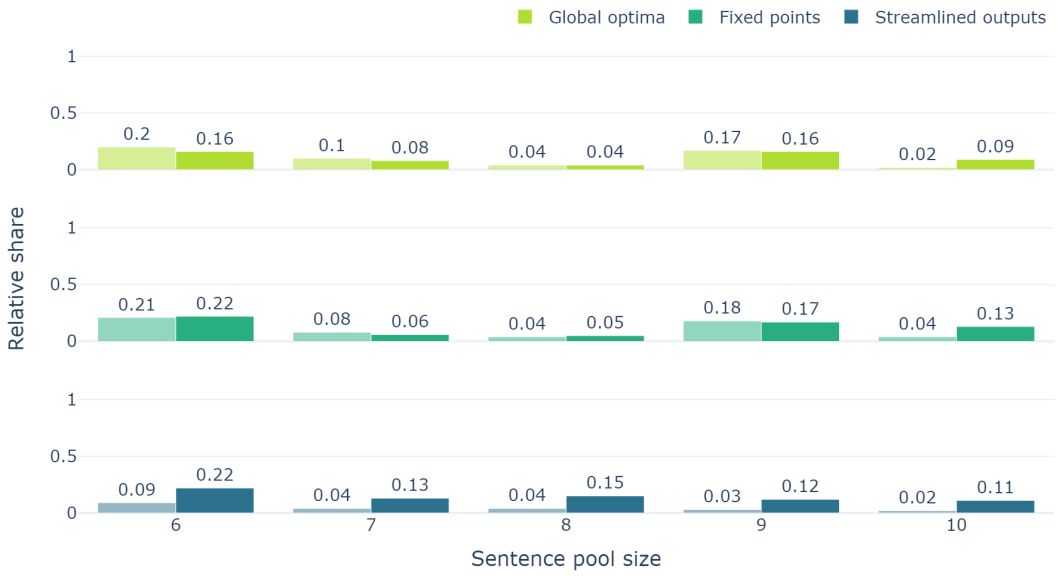
Figure E.3 displays the result for the change in compatibility for initial and output sets of commitments. For the configuration in (A), we can observe a significant boost in compatibility between inputs and outputs. It is most pronounced for global optima, which perform better than the streamlining outputs. For fixed points the results are mixed as the relative share of compatible outputs decreases with the number of involved sentences. For the second configuration in (B), neither global optima nor fixed points manage to improve upon initial compatibility.

Figure E.4 presents the results for plotting the initial similarity among paired sets of initial commitments (horizontal) against the similarity of corresponding paired output sets of commitments. Global optima and fixed points improve upon the dashed diagonals, which signify parity between initial and output similarity, for low (both) and medium (global optima) initial similarity for the first configuration in (A). For the second configuration of weights in (B), global optima and fixed points are very close the dashed diagonals.

By large and far, the results are consistent with the findings presented earlier in the chapter, and they are mostly robust with respect to the small variation of sentence pool size.

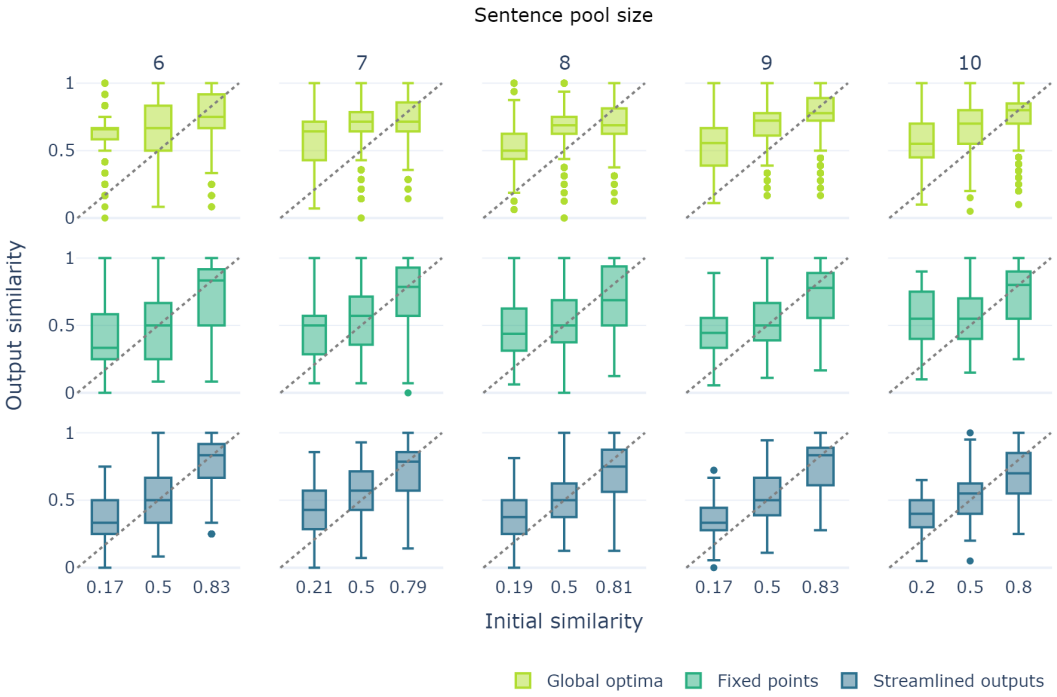


(A) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$

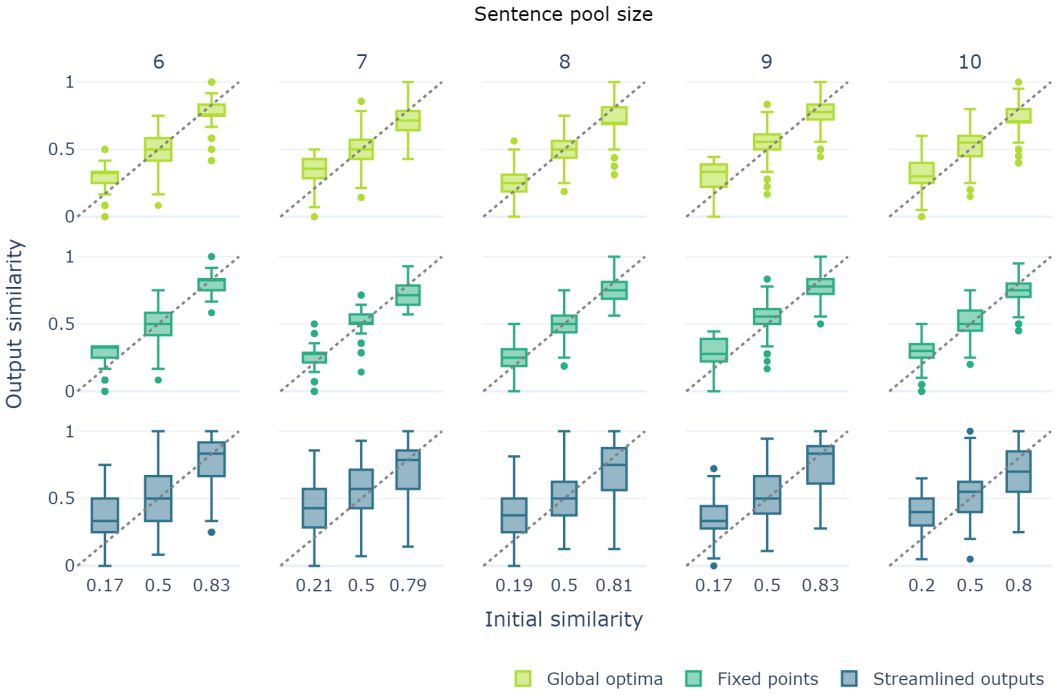


(B) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$

FIGURE E.3: Relative share of compatible pairs of sets of initial commitments (light shade) and pairs of output sets of commitments (dark shade) grouped by sentence pool size.



(A) $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$



(B) $(\alpha_A, \alpha_S, \alpha_F) = (0.10, 0.35, 0.55)$

FIGURE E.4: Plotting output similarity against initial similarity bins (low, medium, high) grouped by sentence pool size. The grey, dashed line indicates parity between input and output similarity.

Chapter 12

Discussion

Theoretical virtues in RE are a vast and under-explored field of research. Running counter to the title of this dissertation – "Virtuously Circular" – I reckon that we did not come full circle, and I take this to be virtuous. In this concluding chapter, I aim to summarise where I have achieved some progress (Section 12.1), and discuss the limitations of the present project with an outlook to further research (Section 12.2).

12.1 Lessons for RE

Recall the research question from Section 1.2:

> How can we integrate theoretical virtues into an account of RE such that they play an active role in addressing objections to the justificatory power of RE?

Here is a summary of the answer that I developed over the course of this project. In view of the issues of ambiguity and trade-offs raised in philosophy of science, an appropriate way to integrate theoretical virtues into RE is to develop a configuration. This means that theoretical virtues need to be selected, specified, weighted and aggregated in view of pragmatic-epistemic objectives pursued in an inquiry with RE. In order to provide a "base" configuration of theoretical virtues that are generally relevant to RE, I focused on the objective of coherence. In a deductive framework I developed an account of coherence based on the virtues of consistency, account, syntactical simplicity and scope. In addition, this revealed interrelations to unifying potential, actual unification and non-ad-hocness.

This resulted in a more substantive notion of coherence that goes beyond mere consistency and the vague idea of everything fitting together. It counteracts the tendency to characterise the state of equilibrium too weakly, which then leads to the suspected weakness of RE as an account of justification.

Moreover, these generally relevant theoretical virtues for RE also help to spell out the often alluded to idea of systematisation, as they mark the difference between a unordered collection of elements and a genuine system.

Next, an inspection of the full-fledged formal model of Beisbart, Betz, and Brun (2021) showed that it implements these virtues. Moreover, the formal model assigns relative weights to the measures for gradual virtues an aggregates them in an additive achievement function. This completes the illustrative configuration of theoretical virtues for RE in view of the objective of coherence.

The integration of theoretical virtues into RE by configuring them allows them to play an active role in addressing the objections. This is illustrated in two simulation studies for the conservativity and no-convergence objections to RE, respectively. For some configuration of weights, the model performs better than a streamlining baseline, which involves theoretical virtuous only sparingly. The model performs better than what objections would lead us to expect with respect to many operationalised aspects. For other configurations, the model's performance is unsatisfactory, and it even falls short of the baseline. This illustrates the importance of weighing theoretical virtues and other desiderata of RE.

Now, we also have the opportunity to look at some points raised by Beisbart, Betz, and Brun (2021) during the discussion of their formal model, e.g., concerning the configuration of weights that guide the trade-offs between desiderata (Beisbart, Betz, and Brun, 2021, 458). At one extreme, one could hope to find a particular configuration of weights or a small range of configurations, for which RE yields plausible results expected by proponents. This cannot be corroborated with the present work. The selection of promising configurations of weights in Section 9.3, and the results in the simulation studies indicate that there are multiple configurations of weights that yield satisfactory behaviour. Consequently, more "fine-tuned" configurations of weights will have to depend on other pragmatic-epistemic objectives or the contexts of specific applications of RE.

At the other extreme, one could claim that agents are completely free to choose the weights as they like. This does not hold in the formal model either. I was also able to identify some basic restrictions that delineate a range of plausible configurations of weights. If faithfulness receives more weight than account, the formal model is unlikely to reach full RE states with equilibration processes (Section 9.3), and it no-longer evades the objections of conservativity (Chapter 10) and no-convergence (Chapter 11). There is also a

plausible trade-off between syntactical simplicity and scope involved in the measure of systematicity (Section 8.1). Increasing the syntactical complexity of a theory is admissible if it leads to a relative increase in scope that exceeds the relative increase in complexity. This is the kind of trade-off which could be expected of theories that exhibit the virtue of unifying potential.

Concerning the more specific trade-off between account and systematicity, there are mixed results. Giving systematicity a lot of weight, e.g., the standard configuration $(\alpha_A, \alpha_S, \alpha_F) = (0.35, 0.55, 0.10)$ used by Beisbart, Betz, and Brun (2021), leads to unsatisfactory performances of the model with respect to the baseline in some aspects, e.g., axiomatic systematicity (Section 10.5), or intrapersonal convergence to a unique output (Section 11.2). This led me to prefer another configuration for the purpose of presentation that trades off account and systematicity differently: $(\alpha_A, \alpha_S, \alpha_F) = (0.55, 0.35, 0.10)$. However, this could also be an artefact introduced by the shortcoming of the measure of systematicity not being able to discriminate singleton theories on the basis of their scope, or due to my preoccupation with output commitments, which profit from high values for account. For model variants that overcome the shortcoming of the systematicity measure[1], one would have to determine promising configurations of weights anew. The considerations of Chapter 9 may serve as a template. I doubt that configurations of weights for small variations of the default model would differ starkly from the present ones.

Thus, I think that the present work recommends to take a position in the middle between a uniquely best configuration of weights and "anything goes". I suppose that it is appropriate to take the same position on the more general level of configuration of theoretical virtues for RE and not just on the level of their weighting. Plausibly, there are multiple ways to configure theoretical virtues for RE, even in view of shared pragmatic-epistemic objectives, but this still does not allows for "anything goes" with respect to configurations. For example, as long as we consider coherence as an objective of inquiry by RE, virtues such as consistency, account, simplicity and scope should be present to render the notion of coherence sufficiently substantial. Moreover, in view of the fact that Proposition 5 establishes that "anything Pareto efficient goes" for gradual virtues, the inclusion of some virtues as necessary requirements is highly recommendable.

Next, Beisbart, Betz, and Brun (2021, 459) state the following:

[1]For example, we could split the measure of systematicity into a measure for syntactical simplicity and a separate measure for scope, and assign them relative weights in the achievement function besides account and faithfulness.

> A strong case for RE can be made if there is a set of weights such that the model behavior has a lot of desirable features, while at the same time evading objections.

I think that I am in a position to present this strong case for RE. There are configuration of weights that exhibit many desirable features, e.g., removing inconsistencies, or reaching full RE states through equilibration, and evade the objections of conservativity and no-convergence effectively. Notably, the same configuration of weights tend to perform well (bad) across all studied aspects. This is a welcome result as widespread conflicts among desirable features would be detrimental to the prospects of providing a plausible configuration.

Even though the results are rather technical in nature due to formalisation and computer simulation, I still take this work to provide enough material to advance the informal discussion about RE as well.

The involvement of theoretical virtues in RE can be traced to the classic accounts of RE, and elaborate bring them to the fore. In contrast, I observe the absence of theoretical virtues in critical stances towards RE (Kappel (2006) being the exception). This allows for proponents of RE to formulate a rejoinder that has not yet been made that pointedly. RE is too weak as an account of justification because the predominant but weak characterisation of coherence as consistency and fit overlooks the involvement of theoretical virtues. The present work provides a proof of concept, illustrating that theoretical virtues can be integrated into RE such that the weakness objections can be dealt with. Theoretical virtues significantly contribute to the justificatory power of RE.

This might break the stalemate that I perceive in the discussion about RE. In addition to informal plausibility considerations, the now available simulation results need also to be taken into account. This bears the potential to shift the discussion to a critical appraisal of the formal model and simulation results.

12.2 Current Limitations and Outlook to Further Research

Formalisation that allows for computer implementation comes at the price of simplification and idealisation. For a discussion of limitations of the formal model, see (Beisbart, Betz, and Brun, 2021, 459–462). I have discussed the limitations of the simulation studies in the respective chapters. Here, I will

add some general points. In many cases, the limitations are an opportunity for further research in attempting to overcome them.

Too Much False Assumptions? Aggregating and trading-off theoretical virtues rest on highly idealised assumptions. Individual orderings of theories according to virtues are assumed to be total, i.e., every pair of theories is comparable. This is extended to the aggregated ordering of overall virtuousness. For weakening of this assumption, in particular an overall ranking of theories that may result in a partial order, see (Priest, 2001). In a next step, one could even look whether there are useful approaches to aggregate partial orderings. However, facing incomparable theories may bring an equilibration process to a premature halting point, or force the agent to proceed tentatively by making an arbitrary choice.[2]

Next, the formal approach allows to define measures on ratio scales, which circumvent incomparability or Arrow's impossibility theorem transferred to theory choice by Okasha (2011) in Section 6.2. However, Rechnitzer (2022, 54) is certainly right in stressing that we better not force everything onto such scales, and revert to ordinal scales. I illustrated in Section 6.1 that there are aggregation rules for such orderings that play out nicely in the illustrative example, even if they fall short of Arrow's requirements. However, the plausibility or relevance of these requirements is open for discussion.

Next, my focus on deductive inference in this project excludes inductive, abductive or probabilistic reasoning, which all are relevant to the idea that inferential relations convey the mutual support of coherence. Hansson (2007) devices a framework with a general support relation to study the compatibility of coherentist and foundationalist requirements. It would be interesting to see, whether RE, as a weakly foundationalist account of justification, can be implemented in this framework as well.

I do not think that the inclusion of additional kinds of inferential relations will render theoretical virtues redundant. For example, explanation, and especially inference to the best explanation are frequently related to explanatory virtues, which comprise more or less the same elements as discussed here under the rubric of theoretical virtues.

A Word of Warning In view of the limitations, I advise to proceed with extreme caution in applying the formal model of RE and its computational

[2]Note that this is different from the choice between equally virtuous theories, which rests on their comparability.

tools as a method of justification to real-life ethical dilemmas, for example. Even more so, if its application goes beyond exploration, for example towards the integration into autonomous systems. In the formal coherence framework of Thagard (2000), which happens to be labelled "reflective equilibrium", Yilmaz, Franco-Watkins, and Kroecker (2017) computationally explore the example of whether a simulated unmanned vehicle should carry out a lethal strike against terrorists. I cannot even begin to imagine the horrors of a dystopian world where autonomous systems base their actions on "justified beliefs" produced by a computer implementation of RE.

The present formal model does not produce indefeasible justification, it will always be relative to the inputs that we provide, and Rechnitzer's application of RE to the justification of a precautionary principle (2022) revealed the extant amount of de-idealisation that is required to apply RE.

Going Local The present studies are a stepping stone to explore new model variants and a touch-stone for upcoming results. The formal model of Beisbart, Betz, and Brun (2021) involves the distinction of globally optimal states according to the achievement function and the semi-global optimisation in the alternating adjustment steps of equilibration processes. We can the observe the impact of reducing the available alternatives in the simulation studies. For example, global optima tend to promote more agreement in terms of compatibility and similarity than fixed points (Section 11.3).

However, semi-global optimisation is still highly idealised. For an example with 20 sentences, an agent has to survey $3^{20} - 1 = 3{,}486{,}784{,}400$ sets of commitments in each commitment adjustment step. We will quickly run out of brain or computer power to process even moderately sized examples.

Hence, it is a worthwhile endeavour to de-idealise the model towards rationally bounded agents that search *locally* for adjustments. Formally, this could amount to a restriction of candidate positions in a neighbourhood of the current position determined by a fixed value for the Hamming distance. If the value is set to 1, the agent can modify at most one sentence at each step. This would model a "piecemeal" process, yet another under-explored idea that originates from (Goodman, 1983(1955)).

Locally optimising processes have an important practical upshot. Much larger examples become computationally feasible as the search no loner grows exponentially (for piecemeal adjustments). But there is also a downside. We have to forfeit global optimisation, and hence the ability to determine whether a fixed point qualifies as a (full) RE state.

One might examine whether, and for which configuration of weights, local optimising RE processes yield globally optimal states in computationally feasible examples. If locally optimising processes are successful in reaching globally optimal states, we might attempt to extrapolate the results to larger examples.

Alternatively, we might move towards a more proceduralist understanding of RE states. This would mean to revise the definition of an RE state, replacing the global optimality condition with a gradual notion of "optimal in a neighbourhood of depth k". Note that the additional conditions (CCT) and (FEA) can be determined in a local setting, and they do not quickly face computational hurdles.

Going Further At the beginning of this project, I set out to investigate the vast under-explored grounds of theoretical virtues in RE. I managed to explore a little, but much more has yet to be discovered. The present work opens the doors to study RE at the nexus of hitherto unrelated approaches. First, I see a connection in the account of virtue-based coherence in the framework of BRT and the model's formal framework of TDS. By investing some assumptions, belief bases can be translated into positions in dialectical structures, and vice versa. Next, my approach towards resolving the issues of aggregating and trading off virtues, revealed an interesting connection to results in formal approaches to welfare economics, or more general, multi-criteria decision analysis. In analysing the formal model, I was even able to draw on results from the distant field of convex geometry. I am optimistic that more thorough research in these directions would yield further insights into RE, and especially for configuring theoretical virtues for RE.

Of course, there is also important informal work that I have to leave for future research. For example, I did not survey or systematise theoretical virtues in philosophy in general.[3] This would be a central piece of clarificatory work if RE is to be a general method of philosophy, and if we want to make the methodological advise of RE to include theoretical virtues more helpful. Next, it would be interesting to compare the present configuration of theoretical virtues to the selections and rankings of virtues that are based on empirical research among philosophers, as well as natural and social scientists (e.g., Schindler, 2022). Another, in my view, interesting line of research concerns the relation of RE and (philosophy of) science. and most of the literature on RE draw on philosophy of science to make a point about RE. But

[3] (Rechnitzer, 2022; Timmons, 2012)

what if we reverse the roles? For example, can scientific inquiry be reconstructed as a process of equilibration?

The present work is a case in point that the methodological advice to use formal and computational methods proves beneficial to philosophical inquiry. And so does the advice to include theoretical virtues in RE. Here, I cannot offer a very wide, or stable state of equilibrium, but the steps taken help to render some initially tenable commitments about the role of theoretical virtues more systematic.

Bibliography

Albert, Hans (1985). *Treatise on Critical Reason*. Princeton University Press.

Alchourrón, Carlos E., Peter Gärdenfors, and David Makinson (1985). "On the Logic of Theory Change: Partial Meet Contraction and Revision Functions". In: *The Journal of Symbolic Logic*, 510–530. DOI: 10.2307/2274239.

Arras, John D. (2007). "The Way We Reason Now: Reflective Equilibrium in Bioethics". In: *The Oxford Handbook of Bioethics*. Ed. by Bonnie Steinbock. Oxford University Press, 46–71. DOI: 10.1093/oxfordhb/9780199562411.003.0003.

Arrow, Kenneth J. (1951). *Social Choice and Individual Values*. Yale University Press.

Arrow, Kenneth J. and Gerard Debreu (1954). "Existence of an Equilibrium for a Competitive Economy". In: *Econometrica*, 265–290.

Baker, Alan (2022). "Simplicity". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2022. Metaphysics Research Lab, Stanford University.

Bartelborth, Thomas (1999). "Coherence and Explanations". In: *Erkenntnis*, 209–224. DOI: 10.1023/A:1005594409663.

Baumberger, Christoph and Georg Brun (2017). "Dimensions of Objectual Understanding". In: *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*. Ed. by Stephen Grimm Christoph Baumberger and Sabine Ammon. Routledge, 165–189.

— (2021). "Reflective Equilibrium and Understanding". In: *Synthese*, 7923–7947. DOI: 10.1007/s11229-020-02556-9.

Beisbart, Claus, Gregor Betz, and Georg Brun (2021). "Making Reflective Equilibrium Precise. A Formal Model". In: *Ergo*, 441–472. DOI: 10.3998/ergo.1152.

Betz, Gregor (2010). *Theorie Dialektischer Strukturen*. Klostermann.

Betz, Gregor (2012). *Debate dynamics: How controversy improves our beliefs*. Springer Science & Business Media.

Bhakthavatsalam, Sindhuja and Nancy Cartwright (2017). "What's so Special About Empirical Adequacy?" In: *European Journal for Philosophy of Science*, 445–465. DOI: 10.1007/s13194-017-0171-7.

Blanshard, Brand (1939). *The Nature of Thought: Volume II*. London, England: Allen & Unwin.

Bonevac, Daniel (2004). "Reflection Without Equilibrium". In: *Journal of Philosophy*, 363–388.

BonJour, Laurence (1985). *The Structure of Empirical Knowledge*. Cambridge, MA, USA: Harvard University Press.

Bovens, Luc and Stephan Hartmann (2003). "Solving the riddle of coherence". In: *Mind*, 601–633. DOI: 10.1093/mind/112.448.601.

Boyd, Stephen P and Lieven Vandenberghe (2004). *Convex optimization*. Cambridge university press.

Brandt, Richard B. (1979). *A Theory of the Good and the Right*. Oxford University Press.

— (1985). "The Concept of Rational Belief". In: *The Monist*, 3–23.

Brun, Georg (2014). "Reflective Equilibrium Without Intuitions?" In: *Ethical Theory and Moral Practice*, 237–252. DOI: 10.1007/s10677-013-9432-5.

— (2020). "Conceptual Re-Engineering: From Explication to Reflective Equilibrium". In: *Synthese*, 925–954. DOI: 10.1007/s11229-017-1596-4.

— (2022). "Re-Engineering Contested Concepts. A Reflective-Equilibrium Approach". In: *Synthese*, 1–29. DOI: 10.1007/s11229-022-03556-7.

Carnap, Rudolf (1950). *Logical Foundations of Probability*. Chicago]University of Chicago Press.

Christensen, David (1994). "Conservatism in Epistemology". In: *Noûs*, 69–89. DOI: 10.2307/2215920.

Cummins, Robert C. (1998). "Reflection on Reflective Equilibrium". In: *Rethinking Intuition*. Ed. by Michael DePaul and William Ramsey. Rowman & Littlefield, 113–128.

Daniels, Norman (1979). "Wide Reflective Equilibrium and Theory Acceptance in Ethics". In: *The Journal of Philosophy*, 256–282.

— (1980). "Reflective Equilibrium and Archimedean Points". In: *Canadian Journal of Philosophy*, 83–103.

— (1996). *Justice and Justification: Reflective Equilibrium in Theory and Practice*. Cambridge University Press.

Das, I. (1999). "A preference ordering among various Pareto optimal alternatives". In: *Structural Optimization*, 30–35. DOI: 10.1007/BF01210689.

de Maagt, Sem (2017). "Reflective Equilibrium and Moral Objectivity". In: *Inquiry: An Interdisciplinary Journal of Philosophy*, 443–465. DOI: 10.1080/0020174X.2016.1175377.

De Regt, Henk (2017). *Understanding scientific understanding*. Oxford University Press.

DePaul, Michael (2006). "Intuitions in Moral Inquiry". In: *The Oxford Handbook of Ethical Theory*. Ed. by David Copp. Oxford University Press, 595–623.

— (2013). "Reflective Equilibrium". In: *International Encyclopedia of Ethics*. Ed. by Hugh LaFollette. John Wiley & Sons, Ltd, 4466—4475. DOI: 10.1002/9781444367072.wbiee197.

Douglas, Heather E. (2009). *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press.

— (2013). "The Value of Cognitive Values". In: *Philosophy of Science*, 796–806. DOI: 10.1086/673716.

Duhem, Pierre (1954). *The Aim and Structure of Physical Theory*. Translated from the french by Philip P. Wiener. Princeton: Princeton University Press.

Dutilh Novaes, Catarina (2020). "Carnapian explication and ameliorative analysis: a systematic comparison". In: *Synthese*, 1011–1034. DOI: 10.1007/s11229-018-1732-9.

Ebertz, Roger P. (1993). "Is Reflective Equilibrium a Coherentist Model?" In: *Canadian Journal of Philosophy*, 193–214. DOI: 10.1080/00455091.1993.10717317.

Elgin, Catherine (1996). *Considered Judgment*. Princeton: New Jersey: Princeton University Press.

Elgin, Catherine (2017). *True Enough*. Cambridge: MIT Press.

Elgin, Catherine and James Van Cleve (2014). "Can Belief Be Justified Through Coherence Alone?" In: *Contemporary Debates in Epistemology*. Ed. by Matthias Steup, John Turri, and Ernest Sosa. Blackwell, 244–273.

Elgin, Catherine Z. (1997). *Between the Absolute and the Arbitrary*. Cornell University Press.

Elliott, Kevin C. and Daniel J. McKaughan (2014). "Nonepistemic Values and the Multiple Goals of Science". In: *Philosophy of Science*, 1–21. DOI: 10.1086/674345.

Engel, Mylan (1992). "Personal and Doxastic Justification in Epistemology". In: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 133–150.

Ewing, Alfred C. (2012(1934)). *Idealism (Routledge Revivals): A Critical Survey*. London, England: Routledge.

Fitelson, Branden (2003). "A probabilistic theory of coherence". In: *Analysis*, 194–199. DOI: 10.1111/1467-8284.00420.

Foley, Richard (1993). *Working Without a Net: A Study of Egocentric Epistemology*. New York: Oxford University Press.

Forster, Malcolm and Elliott Sober (1994). "How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions". In: *British Journal for the Philosophy of Science*, 1–35.

Freivogel, Andreas (2021). "Modelling Reflective Equilibrium with Belief Revision Theory". In: *The Logica Yearbook 2020*. Ed. by Martin Blicha and Igor Sedlár, 65–80.

Friedman, Michael (1974). "Explanation and Scientific Understanding". In: *The Journal of Philosophy*, 5–19.

Goodman, Nelson (1952). "Sense and Certainty". In: *The Philosophical Review*, 160–167.

— (1955). "Axiomatic Measurement of Simplicity". In: *The Journal of Philosophy*, 709–722.

— (1983(1955)). *Fact, Fiction, and Forecast*. Fourth edition. Cambridge, Massachusetts: Harvard University Press.

Goodman, Nelson and Catherine Z. Elgin (1988). *Reconceptions in philosophy and other arts and sciences*. Indianapolis: Hackett Pub. Co.

Grünbaum, Adolf (1976). "Ad Hoc Auxiliary Hypotheses and Falsificationism". In: *The British Journal for the Philosophy of Science*, 329–362.

Hahn, Susanne (2000). *Überlegungsgleichgewicht(e): Prüfung einer Rechtfertigungsmetapher*. Freiburg (Breisgau); München: Alber.

Hansson, Sven Ove (1996). "Knowledge-Level Analysis of Belief Base Operations". In: *Artificial Intelligence*, 215–235. DOI: 10.1016/0004-3702(95)00005-4.

— (1999). *A Textbook of Belief Dynamics. Theory Change and Database Updating*. Dordrecht: Kluwer Academic Publishers.

— (2006). "Coherence in Epistemology and Belief Revision". In: *Philosophical Studies*, 93–108. DOI: 10.1007/s11098-005-4058-7.

— (2007). "The False Dichotomy between Coherentism and Foundationalism". In: *The Journal of Philosophy*, 290–300. DOI: 10.5840/jphil2007104620.

Hansson, Sven Ove and Erik J. Olsson (1999). "Providing Foundations for Coherentism". In: *Erkenntnis*, 243–265. DOI: 10.1023/A:1005510414170.

Hare, R. M. (1973). "Rawls' Theory of Justice–I". In: *The Philosophical Quarterly*, 144–155.

Harman, Gilbert (1986). *Change in View: Principles of Reasoning*. Cambridge, MA, USA: MIT Press.

Haslett, D. W. (1987). "What is Wrong with Reflective Equilibria?" In: *Philosophical Quarterly*, 305–311.

Hempel, Carl G. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: The Free Press.

— (1983). "Valuation and objectivity in science". In: *A Portrait of Twenty-five Years*. Springer, 277–304.

— (1988). "On the Cognitive Status and the Rationale of Scientific Methodology". In: *Poetics Today*, 5–27.

Heron, John (2020). "Set-Theoretic Justification and the Theoretical Virtues". In: *Synthese*, 1245–1267. DOI: 10.1007/s11229-020-02784-z.

Hirsch Hadorn, Gertrude (2018). "On rationales for cognitive values in the assessment of scientific representations". In: *Journal for General Philosophy of Science*, 319–331. DOI: 10.1007/s10838-018-9403-6.

Hoyningen-Huene, Paul (2013). *Systematicity: The nature of science*. Oxford University Press.

Ishizaka, Alessio and Philippe Nemery (2013). *Multi-criteria decision analysis: methods and software*. John Wiley & Sons.

Jones, Karen (2005). "Moral Epistemology". In: *The Oxford Handbook of Contemporary Philosophy*. Ed. by Frank Jackson and Michael Smith. Oxford University Press.

Kappel, K. (2006). "The Meta-Justification of Reflective Equilibrium". In: *Ethical Theory and Moral Practice*, 131–147. DOI: 10.1007/s10677-005-9006-2.

Keas, Michael N. (2018). "Systematizing the Theoretical Virtues". In: *Synthese*, 2761–2793. DOI: 10.1007/s11229-017-1355-6.

Keefe, Rosanna (2000). *Theories of Vagueness*. Cambridge University Press.

Kelly, Thomas and Sarah McGrath (2010). "Is Reflective Equilibrium Enough?" In: *Philosophical Perspectives*, 325–359. DOI: 10.1111/j.1520-8583.2010.00195.x.

Kitcher, Philip (1976). "Explanation, Conjunction, and Unification". In: *The Journal of Philosophy*, 207–212.

— (1989). "Explanatory Unification and the Causal Structure of the World". In: *Scientific Explanation*. Ed. by Philip Kitcher and Wesley Salmon. Minneapolis: University of Minnesota Press, 410–505.

Koppelberg, Dirk (2012). "The Significance of Disagreement in Epistemology". In: *Epistemology: Contexts, Values, Disagreement. Proceedings of the 34th International Ludwig Wittgenstein Symposium*. Frankfurt a.M.: ontos.

Kuhn, Thomas S. (1977). "Objectivity, Value Judgment, and Theory Choice". In: *The Essential Tension: Selected Studies in Scientific Tradition and Change*. University of Chicago Press, 320–39.

Kukla, Andre (1994). "Non-Empirical Theoretical Virtues and the Argument from Underdetermination". In: *Erkenntnis*, 157–170.

Lacey, Hugh (1999). *Is Science Value Free?: Values and Scientific Understanding*. Routledge.

Lakatos, Imre (1978). *The Methodology of Scientific Research Programmes: Volume 1: Philosophical Papers*. Cambridge University Press.

Laudan, Larry (1984). *Science and Values: The Aims of Science and Their Role in Scientific Debate*. University of California Press.

— (2004). "The epistemic, the cognitive, and the social". In: *Science, values, and objectivity*, 14–23.

Leplin, Jarrett (1975). "The Concept of an "Ad Hoc" Hypothesis". In: *Studies in History and Philosophy of Science Part A*, 309–345. DOI: 10.1016/0039-3681(75)90006-0.

Levins, Richard (1966). "The Strategy of Model Building in Population Biology". In: *American Scientist*, 421–431.

Lewis, David (1973). *Counterfactuals*. Blackwell.

— (1983). *Philosophical Papers: Volume I*. Oup Usa.

Lipton, Peter (2004). *Inference to the Best Explanation*. 2nd. Routledge.

List, Christian (2022). "Social Choice Theory". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2022. Metaphysics Research Lab, Stanford University.

Little, Daniel (1984). "Reflective Equilibrium and Justification". In: *Southern Journal of Philosophy*, 373–387.

Longino, Helen E. (1996). "Cognitive and Non-Cognitive Values in Science: Rethinking the Dichotomy". In: *Feminism, Science, and the Philosophy of Science*. Ed. by Lynn Hankinson Nelson and Jack Nelson. Dordrecht: Springer Netherlands, 39–58. DOI: 10.1007/978-94-009-1742-2_3.

Lyons, David (1975). "The Nature and Soundness of the Contract and Coherence Arguments". In: *Reading Rawls*. Ed. by Norman Daniels. New York: Basic Books, 141–167.

Mackonis, Adolfas (2013). "Inference to the Best Explanation, Coherence and Other Explanatory Virtues". In: *Synthese*, 975–995. DOI: https://doi.org/10.1007/s11229-011-0054-y.

McGill, Robert, John W. Tukey, and Wayne A. Larsen (1978). "Variations of Box Plots". In: *The American Statistician*, 12–16.

McGrath, Sarah (2019). *Moral Knowledge*. Oxford University Press.

McMullin, Ernan (2008). "The Virtues of a Good Theory". In: *The Routledge Companion to Philosophy of Science*. Ed. by Martin Curd and Stathis Psillos. Routledge, 498–508.

McPherson, Tristram (2015). "The Methodological Irrelevance of Reflective Equilibrium". In: *The Palgrave Handbook of Philosophical Methods*. Ed. by Chris Daly. Palgrave Macmillan, 652–674. DOI: 10.1057/9781137344557_27.

Mizrahi, Moti (2022). "Theoretical Virtues in Scientific Practice: An Empirical Study". In: *British Journal for the Philosophy of Science*, 879–902. DOI: 10.1086/714790.

Morreau, Michael (2014). "Mr. Fit, Mr. Simplicity and Mr. Scope: From Social Choice to Theory Choice". In: *Erkenntnis*, 1253–1268. DOI: 10.1007/s10670-013-9549-x.

— (2015). "Theory Choice and Social Choice: Kuhn Vindicated". In: *Mind*, 239–262. DOI: 10.1093/mind/fzu176.

Negishi, Takashi (1960). "Welfare Economics and Existence of an Equilibrium for a Competitive Economy". In: *Metroeconomica*, 92–97.

Nielsen, Kai (1982). "Grounding rights and a method of reflective equilibrium". In: *Inquiry*, 277–306.

Okasha, Samir (2011). "Theory Choice and Social Choice: Kuhn Versus Arrow". In: *Mind*, 83–115. DOI: 10.1093/mind/fzr010.

— (2015). "On Arrow's Theorem and Scientific Rationality: Reply to Morreau and Stegenga". In: *Mind*, 279–294. DOI: doi.org/10.1093/mind/fzu177.

Olsson, Erik (2021). "Coherentist Theories of Epistemic Justification". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2021. Metaphysics Research Lab, Stanford University.

Olsson, Erik J. (1999). "'Cohering With'". In: *Erkenntnis*, 273–291. DOI: 10.1023/A:1005530006938.

Patty, J.W. and E.M. Penn (2019). "A defense of Arrow's independence of irrelevant alternatives". In: *Public Choice*, 145–164. DOI: 10.1007/s11127-018-0604-7.

Peirce, C. S. (1877). "The Fixation of Belief". In: *Popular Science Monthly*, 1–15.

Peregrin, Jaroslav and Vladimír Svoboda (2017). *Reflective Equilibrium and the Principles of Logical Analysis: Understanding the Laws of Logic*. Routledge.

Popper, Karl (1959a). "Testability and 'Ad-Hocness' of the Contraction Hypothesis". In: *British Journal for the Philosophy of Science*, 50. DOI: 10.1093/bjps/x.37.50-a.

— (1959b). *The Logic of Scientific Discovery*. Routledge.

Priest, Graham (2001). "Paraconsistent Belief Revision". In: *Theoria*, 214–228. DOI: 10.1111/j.1755-2567.2001.tb00204.x.

— (2016). "Logical Disputes and the a priori". In: *Logique et Analyse*, 347–366.

— (2019). "Logical Theory Choice". In: *The Australasian Journal of Logic*, 283–297. DOI: 10.26686/ajl.v16i7.5917.

Psillos, Stathis (1999). *Scientific Realism: How Science Tracks Truth*. Routledge.

Quine, Willard. V. O. (1951). "Main Trends in Recent Philosophy: Two Dogmas of Empiricism". In: *The Philosophical Review*, 20–43.

— (1955). "Posits and Reality". In: *The Ways of Paradox and Other Essays*. 2nd. Cambridge, MA: Harvard University Press, 246—254.

Quine, Willard. V. O. and Joseph S. Ullian (1978). *The Web of Belief*. Second. New York: Random House.

Rawls, John (1951). "Outline of a Decision Procedure for Ethics". In: *Philosophical Review*, 177–197. DOI: 10.2307/2181696.

— (1971). *A Theory of Justice*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

— (1974). "The Independence of Moral Theory". In: *Proceedings and Addresses of the American Philosophical Association*, 5–22.

— (1980). "Kantian Constructivism in Moral Theory". In: *The Journal of Philosophy*, 515–572.

— (1999a). *A Theory of Justice: Revised Edition*. Cambridge, Massachusetts: Harvard University Press.

— (1999b). *Collected Papers*. Ed. by Samuel Freeman. Harvard University Press.

Rechnitzer, Tanja (2022). *Applying Reflective Equilibrium. Towards the Justification of a Precautionary Principle*. Cham: Springer.

Rott, Hans (2000). "Two Dogmas of Belief Revision". In: *The Journal of Philosophy*, 503–522. DOI: 10.2307/2678489.

— (2001). *Change, Choice and Inference: A Study of Belief Revision and Nonmonotonic Reasoning*. Oxford, England: Oxford University Press.

Sayre-McCord, Geoffrey (1996). "Coherentist Epistemology and Moral Theory". In: *Moral Knowledge? New Readings in Moral Epistemology*. Ed. by Walter Sinnott-Armstrong and Mark Timmons. Oxford University Press.

Scanlon, Thomas M. (2003). "Rawls on Justification". In: *The Cambridge Companion to Rawls*. Ed. by Samuel R. Freeman. Cambridge University Press, 139–167. DOI: 10.1017/CCOL0521651670.004.

— (2014). *Being Realistic About Reasons*. Oxford: Oxford University Press.

Scheffler, Israel (1954). "On Justification and Commitment". In: *The Journal of Philosophy*, 180–190.

Schindler, Samuel (2018). *Theoretical Virtues in Science : Uncovering Reality Through Theory*. Cambridge University Press.

— (2020). "Theoretical Virtues in Science". In: *Oxford Bibliographies in Philosophy*. Ed. by Duncan Pritchard. Oxford University Press. DOI: 10.1093/OBO/9780195396577-0409.

— (2022). "Theoretical Virtues: Do Scientists Think What Philosophers Think They Ought to Think?" In: *Philosophy of Science*, 542–564. DOI: 10.1017/psa.2021.40.

Schurz, Gerhard (1999). "Explanation as Unification". In: *Synthese*, 95–114. DOI: 10.1023/A:1005214721929.

Sen, Amartya (1970). *Collective Choice and Social Welfare*. Holden-Day.

Setiya, Kieran (2012). *Knowing Right From Wrong*. Oxford, GB: Oxford University Press.

Singer, Peter (1974). "Sidgwick and Reflective Equilibrium". In: *The Monist*, 490–517.

— (2005). "Ethics and Intuitions". In: *The Journal of Ethics*, 331–352. DOI: 10.1007/s10892-005-3508-y.

Sober, Elliott (2002). "What is the Problem of Simplicity?" In: *Simplicity, Inference, and Modelling*. Ed. by Arnold Zellner, Hugo A. Keuzenkamp, and

Michael McAleer. Cambridge: Cambridge University Press, 13–32. DOI: 10.1017/CBO9780511493164.002.

Steel, Daniel (2010). "Epistemic Values and the Argument From Inductive Risk". In: *Philosophy of Science*, 14–34. DOI: 10.1086/650206.

Stein, Edward (1996). *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*. Oxford, England: Clarendon Press.

Stich, Stephen P. and Richard E. Nisbett (1980). "Justification and the Psychology of Human Reasoning". In: *Philosophy of Science*, 188–202.

Strong, Carson (2010). "Theoretical and Practical Problems with Wide Reflective Equilibrium in Bioethics". In: *Theoretical Medicine and Bioethics*, 123–140. DOI: 10.1007/s11017-010-9140-2.

Tersman, Folke (1993). *Reflective Equilibrium an Essay in Moral Epistemology*. Coronet Books.

— (2008). "The reliability of moral intuitions: A challenge from neuroscience". In: *Australasian Journal of Philosophy*, 389–405. DOI: 10.1080/00048400802002010.

— (2018). "Recent Work on Reflective Equilibrium and Method in Ethics". In: *Philosophy Compass*, e12493. DOI: 10.1111/phc3.12493.

Thagard, Paul (1978). "The Best Explanation: Criteria for Theory Choice". In: *Journal of Philosophy*, 76–92. DOI: 10.2307/2025686.

— (1988). *Computational Philosophy of Science*. MIT Press.

— (2000). *Coherence in Thought and Action*. MIT Press.

Timmons, Mark (2012). *Moral theory: An introduction*. Rowman & Littlefield Publishers.

Tulodziecki, Dana (2012). "Epistemic Equivalence and Epistemic Incapacitation". In: *British Journal for the Philosophy of Science*, 313–328. DOI: 10.1093/bjps/axr032.

Turri, John, Mark Alfano, and John Greco (2021). "Virtue Epistemology". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2021. Metaphysics Research Lab, Stanford University.

van Fraassen, Bas C. (1980). *The Scientific Image*. Oxford University Press.

Van Rossum, Guido and Fred L. Drake (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Weisberg, Michael (2006). "Robustness Analysis". In: *Philosophy of Science*, 730–742. DOI: 10.1086/518628.

Winther, Rasmus Grønfeldt (2021). "The Structure of Scientific Theories". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2021. Metaphysics Research Lab, Stanford University.

Yilmaz, Levent, Ana Franco-Watkins, and Timothy S Kroecker (2017). "Computational models of ethical decision-making: A coherence-driven reflective equilibrium model". In: *Cognitive Systems Research*, 61–74. DOI: 10.1016/j.cogsys.2017.02.005.