How far does reflective equilibrium take us? Investigating the power of a philosophical method

(C. Beisbart and G. Betz)

1. Summary of the research plan

How powerful is philosophical reflection? Can it promote consensus, when different parties disagree on a philosophical problem? How far does it take non-ideal agents with bounded rationality in justifying their views? In this project, we aim to address such questions by analyzing the power of a popular philosophical method, viz. reflective equilibrium (RE, for short). RE is often appealed to in ethics and other parts of philosophy and has become the center of an intensive debate. Proponents of RE have invested high expectations in it and appealed to it e.g. to underpin a realist understanding of certain domains of discourse. Critics, by contrast, have argued that RE is (i) too implausible to grant justification, (ii) too difficult to apply in practice and (iii) not powerful enough to lead to consensus formation. So far, the debate between proponents and critics of RE suffers from a central shortcoming though: It lacks a common ground as to what exactly RE amounts to. Most characterizations of RE have remained too unspecific and vague to allow much progress.

The aim of the proposed research is a new and much more thorough-going investigation of the power of RE. We wish to determine what it can accomplish and what its limitations are by drawing on a clarification of RE that we have obtained in recent research. In particular, we have available an operationalization of RE and a formal model that can be evaluated using computer simulations.

To apply RE, an agent starts out with her commitments about a specific topic and then puts pressure on them by confronting them with a systematic theory. The fundamental unit of investigation thus is a (dual) epistemic state that consists of a set of commitments and a theory. Our operationalization of RE usefully distinguishes two kinds of aspects within the approach. The *static aspects* comprise the desiderata on epistemic states, viz. systematicity of the theory, its ability to account for the commitments and the faithfulness of the commitments to the initial view. Our model specifies measures that quantify to which extent these desiderata are fulfilled; it further fixes a trade-off between the desiderata. The *dynamical aspects* of the method, by contrast, encompass rules that characterize a dynamic process of equilibration. In previous research, we have developed software in which the rules can be applied in a stepwise manner.

To assess the power of RE, we propose to answer the following research questions:

1. How plausible is RE as a method of justification in philosophy?
2. How practicable is RE, in particular for non-ideal agents?
3. To what extent does RE reduce disagreement between different agents?
4. What are the meta-ethical implications of RE, if it is applied to judgments about reasons?

The first and the second questions address the charges of implausibility (item i above) and impractability (ii), respectively, while the others deal with the worry that RE does not foster consensus formation (iii). The fourth question also refers to a central hope that prominent meta-ethical realists and objectivists have connected with RE.

4 Ph.D. students, directed by the applicants and by a research coordinator, will answer these questions by combining conceptual work with an analysis of our model. Thereby, our previous clarification of RE will not be taken as fixed, but rather be amended if this proves appropriate. We draw on insights from, and expect repercussions for, several recent philosophical debates, in particular about theoretical virtues (1), bounded rationality (2), peer disagreement (3) and realism vs. constructivism regarding practical reasons (4).

2. Research plan

Method matters. The research strategies and modes of investigation chosen for a field do not only provide means to accomplish pre-determined goals of researchers. The repertoire of methods available to a community of scholars also constrains the research questions they can fruitfully pursue. Methods open pathways to new results, but also sometimes close avenues. They determine what counts as result and as achievement. They not only reflect the knowledge of a scientific community, but also the values that members of this community accept. It does not come as a surprise then that controversies in the sciences and humanities often have a methodological side. Nor is it a wonder that philosophers of science have intensely investigated the methods used in the sciences and the humanities (see Kuhn 1962, Lakatos 1970, Feyerabend 1975 for classic examples).

Method matters in philosophy, in particular. Given a traditional understanding of the discipline, it does not draw on data from phenomena, but rather proceeds a priori. Since philosophy lacks the type of empirical basis on which the natural and social sciences are built, the choice of a method becomes even more important than it is in the sciences. Well-known philosophical methods include e.g. conceptual analysis, Carnapian explication, or the methods of phenomenological analysis such as Husserl's method of *epoche*. But many of them are controversial. Conceptual analysis, for instance, has been criticised as being too conservative (e.g. Haslanger 2012), and it has recently come under pressure from the experimental philosophy movement (see e.g. Knobe et al. 2008). The methods constitutive of phenomenology seem too much tied to a specific philosophical outlook as to find much agreement.

A method that is often appealed to in various philosophical circles and that doesn't seem to commit one to a specific philosophical strand is *reflective equilibrium* (RE). Originally introduced by N. Goodman (1955) in philosophy of science, it was later adopted in ethics by Rawls (1971). The method is now often appealed to in normative ethics and related fields. In a recent overview, Tersman (2018) concludes that RE "remains the most popular approach to questions about method in ethics." But the method has also been recommended for other fields, e.g. philosophy of science (Ladyman 2002:54). Some philosophers, e.g. D. Lewis (1983:x-xi), have gone as far as to claim that reflective equilibrium is *the* method of philosophical investigation.

But does reflective equilibrium live up to the hopes that its supporters have invested in it? And can it avoid pitfalls that critics have worried about? The aim of the project is to investigate the power of the method; i.e., to determine what it can accomplish and what its limitations are. As we will show in the next section, a comprehensive assessment of RE is still lacking.

2.1 Current state of research in the field

RE was first proposed by Goodman (1955), when he considered the justification of inferences. Goodman's main idea was that specific inferences, on the one hand, and rules of inference, on the other hand, are justified to the extent that they dovetail with each other. Rawls (1971), to whom we owe the term "reflective equilibrium", claimed to apply the method in developing his theory of justice. Although Rawls's use of the method is intertwined with his contractualism, the following more general picture of RE emerged: An agent starts out with initial commitments about a specific topic, e.g. with a set of moral judgments. The commitments then are explained in terms of a systematic theory, e.g. an ethical theory such as utilitarianism or Rawls's own theory. Once a theory has been identified that best accounts for the commitments, the latter are reconsidered in view of the theory and, if necessary, adjusted to improve fit with the theory. Iterating these steps, the agent moves back and forth between adjusting the theory to commitments and tuning the latter to the former, until an equilibrium state is attained. Such a state is a so-called narrow RE; wide RE additionally takes into account background theories that need to enjoy support from independent

commitments and that are required to speak in favour of the theory to be developed (Rawls 1975, Daniels 1979; see Hahn 2000 on various conceptions of RE). Similar conceptions of RE as the one by Rawls have been considered in moral epistemology (DePaul 2011, Tersman 1993) and applied ethics (Beauchamp & Childress 2009, van der Burg & van Willigenburg 1998, van Thiel & van Delden 2010).

Further elaboration of the method is due to Elgin (1996, 2014), who follows ideas by Goodman (1955, 1977) and Scheffler (1963) to the effect that RE should form the basis of a general epistemology. Elgin argues that our epistemic endeavours do not only strive for knowledge, but also for understanding. Since the latter is advanced by theories that are simple and coherent, theoretical virtues like simplicity and coherence become important in her conception of RE. This is very natural since an influential tradition in philosophy of science (Kuhn 1962, 1977, Hempel 1988) stresses the importance of such virtues for theory choice. Elgin further demands that a reflective equilibrium properly reflects the initial commitments. Baumberger & Brun (2017) condense insights from Elgin and others to give an account of objectual understanding in terms of RE (see also Baumberger, forthcoming).

Since method matters, RE itself has become a matter of controversy.

On the one hand, some authors refer to the method to advance philosophical claims. Michael Smith (1994), for instance, has argued that judgments about what we have reason to do reduce to claims about what more rational selves would desire that we do. The rational advisors are supposed to reason from our own desires and to use RE to make our desires more coherent (159-161). Now clearly, Smith's analysis of judgements about reasons will only lead to similar reasons for different agents, as we expect it for some moral reasons, if the rational advisors converge on the desires by applying RE to different sets of initial desires. Smith's (1994) realism about reasons is thus built upon the expectation that the application of RE leads to consensus formation quite independently from the input various agents start with. This is a very significant expectation that needs further scrutiny (see below for qualifications on Smith's view).

More generally, realist and cognitivist views about practical reasons or morality owe us an account of how we may obtain related knowledge. Some such accounts, e.g. Brink's (1989), Nida-Rümelin's (1996) and Scanlon's (2014), have crucially appealed to RE or similar ideas to propose an epistemology for judgements about practical reasons or moral duties (see Altehenger et al. 2015 for a critical perspective on Scanlon and the contributions in von der Pfordten 2015 for a discussion of Nida-Rümelin).

On the other hand, it has been doubted that RE lives up to high aspirations. The most important strands of criticism can be summarized as follows:

A *first* line of critique takes RE to be an *implausible* account of justification. The most prominent objection in this regard is that every application of the method requires intuitions (Singer 1974, Cummins 1998) or a similar type of input (Stich & Nisbett 1980, Kelly & McGrath 2010). This input then is criticised as being unjustified or unreliable, e.g. with reference to experiments from neuro-science (Singer 2005, see Tersman 2008 for discussion) or from experimental philosophy (see e.g. Sosa 2007, Copp 2012). The charge against RE then is that the results of applying the method are not reliable if the input is not. According to a related objection, RE is implausible because it is too conservative: It only allows for minor adjustments of one's commitments but not for a thorough-going revision, or so it is argued (e.g. Brandt 1985). RE is also taken to be implausible since it is supposed to be trivial (at least in some version, e.g. Singer 2005:347), not an informative description of philosophical scrutiny (Williamson 2007:5, 244-6) or dependent on arbitrary decisions on how to frame a problem (List & Valentini 2016).

A *second* line of criticism holds that RE is *impracticable*. In this vein, Bonevac (2004) doubts that the recommended process of equilibration leads to an equilibrium state (see also List & Valentini 2016 for this worry). A related objection is that

an agent cannot know whether or not she has reached an RE state (Strong 2010). And even a defender of RE, DePaul (2011), thinks that there is a serious problem about RE because it is too demanding.

The *third* line of criticism addresses the expectations that objectivists and realists have invested in the method (see above for examples). The gist of the criticism is that RE is not powerful enough to underwrite a realist or objectivist reading of a domain of discourse. A standard objection in this vein is that RE does not produce consensus when people apply the method starting from different initial conditions (e.g. de Maagt 2017). This leads to the complaint that RE alone cannot grant the objectivity of morality (e.g. de Maagt 2017; see Holmgren 1987 for discussion) and that it cannot explain moral knowledge (McPherson 2015). There are also other reasons to doubt that a domain of discourse in which claims are defended using RE allows for a realist reading. For instance, at least in some versions, RE includes theoretical virtues that seem pragmatic only, e.g. simplicity (Kappel 2006, Kelly & McGrath 2010). Relatedly, RE is supposed to commit one to coherentism about justification, which then is rejected, e.g. because mere coherence cannot be shown to be truth-conducive (e.g. Thagard 1988; see Lycan 2011 for a discussion).

Proponents of RE have of course tried to refute such objections. To give just a few examples regarding the first strand of criticism, Brun (2014) has argued that RE does not rely on intuitions as input (see Knight 2006 for related points), while Daniels (1979), Brink (1989) and DePaul (2011) have claimed that conservativeness can be avoided, in particular if wide RE is practised. So there is a lively and controversial debate about the power of RE (see Walden 2013 and Tersman 2018 for recent defences of RE).

Unfortunately though, the discussion about RE is significantly hampered by a lack of clarity about what exactly RE is. Although Daniels, Elgin and a few others (e.g. Tersman 1993, DePaul 2011) have developed more detailed accounts of RE, the characterization of RE has remained quite vague until very recently. For instance, in the literature, we find almost no formulations for rules which an agent may follow to reach equilibrium. Likewise, it has been unclear what exactly qualifies as initial commitment. As a consequence, many philosophers seem to take RE to be a mere metaphor. A particularly important theoretical issue that has barely been noticed is this: Typical descriptions of RE feature *static* and *dynamic* aspects. Paradigmatically, Goodman (1955) refers to a desired *state* of "agreement" and a *process* of "adjustments." But often these two aspects of RE are not at all distinguished. The question then is how exactly the desired states and the process are related and which has normative primacy. Also, discussions of RE so far suffer from the shortcoming that, although the method of RE is regularly claimed to be employed, there are no studies in which the method is applied step by step in an explicit and transparent way (cf. Doorn 2009). Studies that go some way in this direction, e.g. Swanton's (1992), Hahn's (2000) or some of the contributions in van der Burg & van Willigenburg (1998) do not really give an account of a step-by-step execution of a RE-process for a realistically complex problem, or move in directions that are interesting, but in our view too far from the original descriptions of RE (e.g. van Thiel & van Delden 2010).

It is clear that we cannot reliably assess the power of RE if the method is not well described. Rawls himself pointed out as early as 1971 that "reflective equilibrium suggests straightway a number of further questions. For example, does a reflective equilibrium (in the sense of the philosophical ideal) exist? If so, is it unique? Even if it is unique, can it be reached? Perhaps the judgments from which we begin, or the course of reflection itself (or both), affect the resting point, if any, that we eventually achieve." (Rawls 1971:50) At the time, he noted that "[i]t would be useless, however, to speculate about these matters here. They are far beyond our reach." (ibid.) It is an embarrassment that none of these questions have been studied in a rigorous manner up to now. The reason is that no precise characterization of RE has been available.

So we need, first of all, a clarification of what the method itself is. Our previous research has taken decisive steps in this direction. We therefore propose to draw on this clarification to investigate the power of RE in a more precise way.

2.2 Current state of own research in the field

In a research project "**Das Überlegungsgleichgewicht – Neukonzeption und Anwendung (Reflective equilibrium – re-conception and application)**" funded by the SNSF (no. 150251, 1.9.2014 – 31.8.2018), one applicant, Claus Beisbart (CB), and Georg Brun (GBr) have collaborated to clarify the method of RE. In the course of this project, it turned out to be useful to join forces with the other applicant for this new project, Gregor Betz (GBe). In what follows we describe our results. For simplicity, we'll focus on narrow RE and thus neglect background theories. This is not a problem because the model is easily extended to wide RE. It is in fact our intention to consider both narrow and wide RE in our research.

In a first step, we have obtained an *informal clarification* of RE. Taking Elgin's work (1996, 2014) as a starting point and building on our previous research (Brun 2009, 2012, 2014), we have provided a comprehensive analysis of the various elements of RE-states and processes, as well as an analysis of how RE fits into an account of understanding (Baumberger & Brun 2017). We have also clarified the relation between RE and another important philosophical method, viz. Carnapian explication (Brun 2017a). This research revealed that the notion of equilibrium allows for two related readings: equilibrium as an agreement between commitments and a theory, and equilibrium as a balance between two opposite forces, viz. of respecting initial commitments and of doing justice to desiderata such as simplicity. But it has also become clear that there are limits to what can be accomplished by an informal clarification of the idea of a RE. Real progress towards a fully and precisely articulated conception of RE calls for the help of formal methods.

We have thus proposed (i) an *operationalization* as well as (ii) a *formal model* of RE by clarifying its main components, their relationships, the standards constitutive of RE, and the rules that define a RE process (Beisbart, Betz & Brun under review, see attachment). Both the operationalization and the model are summarized in Table 1 and will be described in what follows.

The guiding idea behind our operationalization is that a single epistemic agent starts out with her commitments about a certain topic and then puts them under pressure by confronting them with theories that try to systematize them. So, the main components of RE are *commitments* and *theories* (plus background theories for wide RE). We characterize commitments and theories merely in terms of their functional roles. Neither content nor form of the initial commitments are restricted in any way: They need neither be consistent nor consist of intuitions nor refer to particular cases. In this way, influential criticism against RE can be avoided upfront (see Brun 2014, with reference to e.g. Cappelen 2012). What sets commitments apart from theories is rather the characteristic role they play in the RE, viz. that they need to be accounted for by theories. The latter are assumed to be consistent. A set of commitments and a theory form what we call a *(dual) epistemic state*.

To clearly separate the static and dynamic aspects of RE (see Sec. 2.1 above), our operationalization first identifies standards that apply to epistemic states and that are independent of any rules for how to proceed. Proposals for rules can then be assessed in terms in terms of the probability with which they lead to a state in which the desiderata are fulfilled to a high degree.

We identify three *desiderata* that are constitutive for RE: first, the ability of the theory to *account* for the commitments, second, the *systematicity* of the theory and, third, the *faithfulness* to the initial commitments. While the first desideratum ties commitments and theories together, the second and third ones each work on one component of the dual epistemic state only and typically pull in different directions: Systematicity acts on the theory and pushes the agent to change her view, whereas

faithfulness acts on the commitments and motivates her to stick to her initial commitments. A central lesson of our operationalization is that we need to define a sensible *trade-off* between the various desiderata constitutive of RE, e.g. in terms of weights.

Once a trade-off has been set or (as we call it) an achievement function has been determined, we can define *optimal* epistemic states: They simply do best in terms of the achievement function. Now even epistemic states that optimize the achievement function may include inconsistent commitments (for instance, if the agent starts with inconsistent commitments and if the desideratum of faithfulness has predominant weight). Thus, such an optimum is only called *full RE state* if the commitments are fully accounted for by the theory, which implies consistent commitments.

Prominent characterizations of RE from the literature feature a process of equilibration, which we explicate as a rule-governed dynamics. In accordance with this literature, we distinguish two types of steps in a RE process. In the first one, the commitments are held fixed and the theory is adjusted to the commitments. In our operationalization of RE, it is natural to conceptualize this as a restricted optimization: The achievement function is optimized while the commitments are held fixed. In the second type of step, the theory is kept constant and the achievement function is optimized by adjusting commitments. We let this process stop as soon as any application of the steps does not lead us any further and a fixed point is reached. Note that the fixed point need not coincide with a full RE state, as defined above. This opens the possibility of a mismatch between what may be called the static and the dynamic aspects of RE. In our proposed research, we plan to have a closer look at this potential mismatch.

Operationalization of RE	**Model of RE**
Topic contemplated by an agent	**Dialectical structure**, i.e., a set of inferentially related sentences
(Dual) epistemic state of the agent, consisting in her *commitments* and a *theory*	Two **sets of sentences** from the dialectical structure that represent the agent's commitments and her theory, respectively
Three desiderata for evaluating an epistemic state, namely: • The theory *accounts* for the commitments; • The theory is *systematic*; • The current commitments are *faithful* to the initial commitments.	**Three measures** that quantify the degree to which the desiderata are fulfilled: • *Account*, depends on the (Hamming) distance between the sets representing commitments and the consequences of the theory; • *Systematicity*, depends on the numbers of principles in the theory; • *Faithfulness*, depends on the (Hamming) distance between the sets representing current and initial commitments.
Trade-off between desiderata: **Achievement function**	**Weighted sum of the measures**
Full RE state, i.e., epistemic state which optimizes the achievement function and in which the commitments are consistent and fully accounted for by the theory.	**Global optimum**, two sets of sentences that jointly maximize the weighted sum of measures and in which the set of commitments is consistent and fully entailed by the theory.
RE process, or **equilibration process**: the rule-governed iterative adjustment of commitments and theories, which is informed by the epistemic desiderata. It may result in a *fixed point*.	**Local optimization**, i.e., iterative restricted optimization of the theory, given the commitments, or the other way around.

TABLE 1: Our operationalization of RE and the formal model.

In our previous research, we have fleshed out our informal operationalization of RE as a *formal, quantitative model*, which was developed in the framework of the *Theory of Dialectical Structures* (Betz 2010, 2012). The model contains formal representations of dual epistemic states, represents the constitutive epistemic desiderata as numerical functions, and allows one to run computer simulations of RE processes. Using stylized examples, we have shown that simulated sequences of states are naturally interpreted as a process of equilibration. We could further show that, in many examples, the process leads to a full RE, although this is not always so. All in all, our model has achieved an important task: It has shown that the ideas underlying RE can be made precise in a consistent manner.

In our previous research, we could also show that our findings about RE have interesting consequences for recent philosophical debates. Brun (2017b) analyses ethical thought experiments from the point of view of RE. He criticizes the widespread view that thought experiments can provide counter-examples that suffice to refute a philosophical theory. Brun (2018) shows how RE can be used to overcome methodological deficits in available accounts of logical expressivism.

But there is still a lot of work to be done. First, our model must be analysed in a more systematic and comprehensive manner. For each topic (modelled as a dialectical structure) and each set of initial commitments, we may ask: Does a full RE state exist? Will it be reached in an equilibration process? And how long does the process take? Second, the results need to be carefully interpreted. A model is at best a simplified counterpart to some part of reality, so the question is which of our findings about the model provide genuine insights about RE. Third, if the model turns out to be unrealistic in some respects, it needs to be de-idealized. In fact, the theoretical framework on which we draw already suggests de-idealizations. It is, for instance, straight-forward to assign weights to the commitments to capture the fact that agents stick to their commitments to different degrees. Note though that such a change of the model requires a substantial amount of technical work.

All in all, from the perspective of our previous clarifications of RE, there is an urgent need for more research. This wouldn't be too significant on itself, if it didn't turn out that the research questions raised by our previous research are closely related to lacunae in the philosophical debate about RE. In particular, a close analysis of our model can help us understand how far RE can take us, e.g. whether it allows for consensus formation.

The need for more research is also manifest from the case study included in the previous project and at the centre of the dissertation of its Ph.D. Tanja Rechnitzer. The case study applies RE to identify a defensible Precautionary Principle that provides moral guidance in situations in which there is the risk of severe harm, for instance with respect to climate change (see e.g. Sandin 1999, Gardiner 2006, Steel 2015 for Precautionary Principles). In the case study, we could for the first time show in detail that and how the crucial ideas behind RE can be applied in practice. But the study has also pointed to challenges for the application of RE. For instance, it has shown that RE requires the agent to clearly decide how much weight the desiderata constitutive of RE should have. Existing accounts of RE remain silent on how exactly the trade-off between the desiderata should be set, so more research is needed at this point (see sub-project 1 below).

Our previous research on RE well qualifies us for the proposed project. The project leaders also have a lot of experience with new aspects of the new project. GBe has developed the Theory of Dialectical Structures (Betz 2010). He has used is to reconstruct debates, to study, e.g., consensus formation and truth-tracking ability (Betz 2012). He has further programmed the computer code that implements our model of RE.

As far as the meta-ethical implications of RE are concerned, CB did his Ph.D. in meta-ethics (Beisbart 2007). He has also done work about theoretical virtues and taught a seminar about them with GBr.

The applicants and the collaborator GBr have a strong track record in collaboration, which includes co-organized workshops (on understanding, organized by GBr, CB et al., Berne 2014; on formal accounts of RE with GBr, GBe and CB,

Berne 2017; on precaution, organized by GBr, CB and T. Rechnitzer, Berne 2017), co-authored papers (see e.g. Brun & Betz 2016; Baumberger, Beisbart & Brun 2017; Beisbart, Betz & Brun under review) and team-teaching in various combinations. GBr also visited GrB at several occasions as an academic guest.

2.3 Detailed research plan

The overarching aim of the proposed project is to investigate the power of RE. We want to obtain the picture of an inquiry that follows the method of RE: How is it guided by theoretical virtues? Will the participants of the inquiry be able to make progress? Will they reach consensus?

In more details, we address the following four research questions that take up hopes and fears associated with RE.

Box RQ. Research questions

1. In how far can RE be understood as a plausible trade-off between theoretical virtues that provides justification f philosophical views?
2. How far does a RE-process enable non-ideal agents with bounded rationality to make progress on their views?
3. Under which conditions, and to what extent, does the application of the method by different agents lead to consens formation?
4. Which meta-ethical view is most appropriate if judgements about practical reasons can be justified using RE?

The first question addresses the first strand of criticism identified above, viz. that RE is implausible. We propose to investigate the theoretical virtues implicit in RE and their trade-offs. The second question reacts to the second strand of criticism and addresses the progress that non-ideal agents can make by applying the method. The other two research questions, finally, deal with the third strand of criticism. They investigate consensus formation and the interpretation of a domain of discourse in which statements are justified using RE. Note that the fourth question also addresses hopes that realist and cognitivist authors have invested in RE.

Each research question is investigated by one dissertation, which is at the centre of a sub-project. In what follows, we will explain the research questions and the related methods. It is important to note, however, that we expect each Ph.D. students to develop her own agenda. The details specified in the following descriptions are therefore meant to point out possible avenues of research; they are not supposed to define tasks that must be carried out completely or exactly as specified here.

Sub-project 1: A plausible method of justification? Trade-offs between theoretical virtues in RE

RE is supposed to justify philosophical views. But does it define a plausible ideal of justification, one that we should strive for? This is, very roughly, the central question of the first dissertation.

Since RE incorporates a theory as a crucial component, it is useful to draw on the discussion about theory choice in philosophy of science. The standard view there is that the choice of theories in the natural and social sciences is guided by a number of *theoretical virtue*s (sometimes also called *epistemic values*), e.g. consistency, empirical adequacy, simplicity, scope and explanatory power (see e.g. Kuhn 1977, Hempel 1988, Lacey 1999; cf. also Lewis 1973:73-77). We can thus say that the theoretical virtues help to justify theories. Since the virtues may pull in different directions, we need a trade-off between the theoretical virtues to compare theories regarding their justification overall.

Likewise, we can assume that the justification of philosophical views depends on the extent to which they fulfil desiderata that have an analogous function as the theoretical virtues. And indeed, desiderata such as systematicity are built into RE: It values epistemic states in which the commitments are well systematized by a theory, but still reflect the initial commitments of the agent. The desiderata are further balanced against each other. This fact is clearly reflected in our model: We have defined an

achievement function that is spelled out as a weighted sum of measures that quantify the degree to which three desiderata are realized. So we can rephrase the central question of the first dissertation as follows: Does RE incorporate a plausible trade-off between relevant desiderata?

Now in our model, the weights for the desiderata constitutive of RE are free parameters. There is thus some leeway in defining what exactly the goal is. So, more precisely, the central question of this sub-project is whether this leeway may be used to determine a sensible trade-off (or a plausible weighting scheme) that articulates how justified a view is overall. This question has two aspects: A) the selection and specification of the desiderata; B) the way they are weighted or aggregated.

A) Selection and specification of desiderata. Do the desiderata implicit in RE plausibly add to justification? And are no important desiderata missing? To answer these questions, we first propose to match the desiderata implicit in RE to theoretical virtues investigated in philosophy of science. This is not to neglect crucial differences between philosophical views and empirical theories. The idea is rather that intensive work in philosophy of science has resulted in a collection of theoretical virtues that provide a useful reference point for a discussion of philosophical theories. We'll thus address the following questions: Do the desiderata implicit in RE have counterparts among the theoretical virtues for scientific theories? If not, is there a plausible explanation for this and can we nevertheless vindicate the desiderata? Second, are there plausible theoretical virtues in the philosophy of science literature that are not included in the desiderata implicit in RE? If not, is there a plausible explanation for this?

We second plan to carry out a little case study to establish an explicit link to *philosophical* theories. The aim is to uncover the desiderata that authors appeal to when arguing about moral theories – be it explicitly or implicitly. The case study will focus on Scanlon's (1998) contractualist theory of what humans owe to each other and recent discussion of it (e.g. Parfit 2011). We will use the case study to answer the same types of questions that we have raised concerning well-known theoretical virtues: Are the desiderata constitutive of RE important in the discussion, and are other desiderata figuring in the debate? In this way, we hope to be able to make a strong case for the desiderata implicit in RE.

In this context, we'll also discuss recent elaborations of RE that include epistemic-pragmatic goals as an additional component. The idea is that, for each specific application of RE, we have to identify more specific goals that constrain the equilibrium too (Elgin 1996, Baumberger & Brun 2017). But this move raises new questions: Why are some goals taken to be constitutive of RE, while others are only optional? Second, what is distinctive of RE if there if there is freedom to consider all kinds of goals?

B) Weighting and aggregation of desiderata. Can the desiderata implicit in RE be balanced against each other in a way that underwrites justification? To answer this question, we may again draw on discussions from philosophy of science, where trade-offs between the theoretical virtues are intensely debated. Kuhn (1977) stresses that these virtues need to be interpreted and weighed against each other and that this cannot be done in terms of an algorithm. Recently, Okasha (2011) has tentatively suggested a parallel between Arrow's theorem and trade-offs between theoretical virtues. If the parallel goes through, we are faced with the disturbing conclusion that no rational trade-off between the theoretical virtues is possible (see Morreau 2013 and Stegenga 2013 for discussion). A more optimistic outlook is implicit in D. Lewis's best system account of laws of nature (Lewis 1973:73-77, see also Lewis 1994). Lewis does not specify a method to strike a balance between the virtues relevant for theories, but he suggests that "nature is kind" because a system is robustly best – that is, a certain theory turns out to be best quite independently of how exactly the weights are set (Lewis 1994:479).

To determine whether, and how, the desiderata implicit in RE may be weighed in a plausible way, we first try to draw on the literature from philosophy of science by addressing the following questions: Are there any proposals for how to justify a

specific trade-off quite generally (see e.g. Huber 2008)? Can these proposals be adapted to RE? How can we account for Kuhn's worries that an algorithmic approach is not possible? In particular, does our model of RE imply an algorithm that is not feasible according to Kuhn? Or should Kuhn's point be taken into account by claiming that the weights cannot be set in a context-independent way? Finally, how can we meet the challenge formulated by Okasha (2011) from the perspective of RE? That is, how can we argue that RE includes a *reasonable* trade-off?

A second line of research is opened up by our previous research: Our operationalization and our model of RE suggest precise conditions which a reasonable trade-off (weights in the achievement function in our model) should fulfil:

a. The trade-off (or the achievement function) should be such that its optima (i.e., epistemic states that score best with respect to it) are full RE states. This implies that the goal of RE guarantees consistency, which is a very basic demand.
b. Optima with respect to the achievement function should be robust; i.e., small changes in the weights should not lead to completely different optimal states (cf. Lewis's optimism about laws of nature, cf. also Rawls 1971:456–7).
c. The trade-off should lead to unique optima; that is, there is just one epistemic state that scores best in terms of the corresponding achievement function. This idea is admittedly more controversial, Whereas Goodman (1955) and Elgin (1996) welcome the possibility that RE can lead to alternative states which are equally in equilibrium, others worry that such pluralism may lead to arbitrariness (e.g. List & Valentini 2016).

The Ph.D. candidate can run computer simulations that implement our model to find out whether there are weighting schemes (parameterizations of the achievement function) that fulfil the conditions (a) – (c) for a broad range of topics (dialectical structures) and various sets of initial commitments (initial conditions of sentences in the model). The results will be related to the more theoretical results obtained from the comparison with the philosophy of science literature.

Box 1: Suggested simulation experiments – sub-project 1:

Step 1. Generate a broad range of settings by varying
- the dialectical structure (e.g., in terms of size),
- the initial commitments.

Step 2. Generate a broad range of achievement functions by varying the
- weights of the measures for the desiderata

Step 3. For each set of initial and boundary conditions and each achievement function determine global optima and test them for consistency, robustness and uniqueness.

We expect two major theoretical payoffs for a further development of our model. First, we hope for the identification of a range of trade-offs that fulfil the conditions mentioned above. The weights constitutive of these trade-offs will be used for research in the other sub-projects. Second, we expect concrete proposals for possible improvements of the way in which we measure the extent to which the desiderata are fulfilled in the model. The desideratum of systematicity is a case in point. So far, in our model, systematicity is only understood as simplicity, which in turn is measured using only the number of axioms a theory consists of (see Table 1). In the sub-project, we will explore additional ideas how to measure systematicity. Fruitfulness, for example, may be measured in terms of the commitments which are taken on board during the RE-process because they follow from the theory.

Note that this sub-project is restricted to what we called the static aspects of RE because the process of equilibration does not play any role. In the second sub-project we turn to the dynamics of the RE-process.

Sub-project 2: A progress of commitments? On the process of equilibration under realistic circumstances

As commonly understood, RE does not only articulate what justified epistemic states are. It also proposes a pathway to obtain (more) justification. Prominent descriptions of RE describe a process that goes back and forth between theories and commitments and mutually adjusts them to each other. But where does this process terminate as a matter of fact? Is the process feasible for non-ideal agents? If not, can we propose alternative rules of thumb that help non-ideal agents make progress on their views? These are the central questions of sub-project 2.

Note that sub-project 2 can be understood as a consistency check for RE. Once we have clearly distinguished between the static and the dynamic aspects of RE (see Sec. 2.1 above), we face the crucial question of whether both fit together. This question has hardly been mentioned in the literature – and, indeed, it is hard to see how it could be addressed without first giving a more specific account of the process of equilibration and the goal it is supposed to achieve. Hence, we expect that real progress can be made on the basis of our formal model.

We propose to systematically study the equilibration process by scrutinizing the model as follows: Starting from initial commitments, we will run simulations, in which both types of steps are carried out until a fixed point is reached. We can then investigate the fixed point: Does it optimize the achievement function? Is it consistent? Is it a full RE state? To what extent are the desiderata constitutive of RE fulfilled? (cf. Box 2).

Of course, we do not expect that the process always runs into an optimum of the achievement function. The equilibration process is defined in terms of restricted optimization problems, whereas the optimum of the achievement function is based upon an unrestricted optimization. We'll look for simple examples in which the process does not yield an optimum and try to understand what the reason is. To assess the effectiveness of the equilibration process, we'll define a success rate that is based upon a close investigation of a large number of dialectical structures and initial conditions (i.e., initial commitments).

In this investigation, the following aspects will receive special attention:

i. Some descriptions of RE (e.g. DePaul 2011; Spohn 2002) imply that it may be difficult to reach RE, and thereby suggest that equilibration takes long. In the simple examples considered so far (Beisbart, Betz & Brun under review), by contrast, a fixed point is often reached after two steps. We'll therefore study the speed with which a fixed point is reached and compare our findings with claims in the literature.

ii. It is possible that, during the process of equilibration, a step doesn't lead to a unique result. For instance, there may be two equally systematic theories that both do equally well in accounting for the commitments. So far, under these circumstances, our model randomly picks one of the optimal theories. But this is problematic, because it may then be more likely that the equilibration runs into a non-optimal state. We'll thus check whether we can do better than picking in a random manner, for example, by branching the process and generating a search tree, where equally good options are pursued in parallel.

All this can be done with our model, but of course the model cannot be expected to reflect reality in every respect. It will therefore be important to check carefully whether the results from the model are stable and whether they reflect the spirit of RE as it has been characterized in the literature. The results will enable us to address some of the objections implicit in the second strand of criticism identified above, viz. that RE is impracticable. In particular, we can discuss doubts to the effect that the RE-process doesn't lead to an equilibrium and that agents cannot know to have reached a RE state.

To address the criticism that the application of RE is too demanding, we propose to explore alternative rules which are tailored to agents with bounded rationality. We may think of them as simple rules of thumb the application of which does not

take many resources. Such rules of thumb will not always significantly improve the epistemic state of the agent; rather, the idea is that the rules are likely to help the agent to make some progress.

To identify sensible candidates for such rules, we can draw on a rich literature on bounded rationality (see e.g. Gigerenzer and Todd 1999, Simon 2000). We can distil the most important restrictions and propose rules that are consistent with them, but also fit the known descriptions of equilibration. In particular, we plan to look at a more piecemeal procedure, for example, rules after which the agent changes at maximum one principle of the theory/one commitment per step. Such a piecemeal strategy is in fact suggested by Goodman's (1955) classic description of RE. Another approach motivated by the philosophical literature is satisficing (e.g. Slote 2001). The idea is that agents with bounded rationality do not have the resources to solve optimization problems, but rather try to reach a certain level of quality. Yet another perspective comes from the literature about optimization problems in computer science (e.g. Lange 2004, Brinkhuis & Tikhomirov 2005). We'll check simple proposals, e.g. the method of steepest ascent, for their ability to facilitate equilibration.

Once we have identified reasonable candidates for rules of thumb, we'll implement them in our model and test them using computer simulations. We can proceed as before with the rules of the original model: We apply the rules of thumb in computer simulations and test whether, and how fast, they lead to a fixed point and how well the epistemic state at the fixed point does in terms of our desiderata.

On the basis of this work, we plan to assess the criticism that RE is impracticable. If we can identify rules of thumb which roughly resemble the common characterizations of the equilibration process, which are easily applied in practice, and which lead to reasonable progress in terms of the desiderata, we can reject the criticism.

Box 2: Suggested simulation experiments – sub-project 2:
Step 1. Generate a broad range of settings by varying
- the underlying dialectical structure,
- the initial commitments.

Step 2. Simulate the equilibration process and analyse
- whether, and how quickly, a fix-point is reached,
- whether the fix-point is a full RE state,
- whether the process itself and the fix-point reached are underdetermined.

Both steps are executed with the rules/dynamics from our current model and rules of thumb suggested by the literature.

Sub-project 3: Agreement secured or disagreement stabilized? RE and consensus formation

As mentioned above in Sec. 2.1, some philosophers have hoped that RE promotes intersubjective agreement. Critics, by contrast, have feared that RE does not lead to consensus (e.g. de Maagt 2017). The worry is, for example, that two agents who each go through a RE process by starting with different initial commitments still end up with different views. Perhaps, it may even happen that disagreement is strengthened when different agents apply RE. Due to the lack of exercisable elaborations of RE, the hopes and worries about consensus formation have so far only been based on qualitative reasoning. The aim of this sub-project is to clarify to what extent, and under which conditions, RE leads to consensus and to trace the philosophical consequences.

As a first step, it is useful to systematically analyse how disagreement may persist, or even arise in the framework of RE. We thus plan to propose a classification of possible sources of disagreement. We assume that two agents start out thinking about roughly the same topic and that each one reaches a fixed point by applying the rules of RE. If there are disagreements on the final commitments, they may be explained in the following ways:

i. The agents differ on their trade-offs because they weigh the desiderata such as systematicity and faithfulness to initial commitments in different ways; for instance, one agent puts more weight on systematicity than another. As a result, we obtain a Kuhnian sort of underdetermination (cf. Kuhn 1977).

ii. The sets of initial commitments differ either because the agents understand the topic in slightly different ways (and thus start with different dialectical structures based upon different sentence pools in our model), or because they have initially different views about exactly the same topic.

iii. In so-called wide RE, the agents differ on their background theories.

iv. In trying to follow the rules of equilibration, at least one agent makes a mistake that the other doesn't commit.

v. At some stage during the application of the method, two or more epistemic states tie for being best, and the agents pick different ones to continue.

This is a list of *possible* sources of disagreement, but the hope is of course that some of them are not significant since the method tends to avoid them. For instance, if RE is *robust* with respect to trade-offs, slight modifications of the trade-offs do not make a difference for the results (which is to say that item i. on the list does not much matter). If the pressure of systematization underlying the method is sufficiently *powerful*, then differences in the initial commitments are washed out and the second source does not have a large impact (cf. ii.). Finally, if the method is *fault-tolerant*, then minor flaws in the application of RE do not impact on the result (cf. iv.).

It is thus useful to investigate in a second step to which extent the application of the method does in fact wipe out the various sources of disagreement. This research question can be answered by studying the behaviour of our model using simulations: We consider pairs of agents who follow the rules of RE under slightly different conditions, e.g. with different sets of initial commitments. When both have reached a fixed point, we compare whether, and to what extent the epistemic states that have been reached differ. The results will be analysed in multiple ways: Under which conditions is consensus formation likely? If no consensus emerges, what are the reasons? Even if there is no consensus, can we at least say that the commitments of both agents have somehow converged? Or is it possible that the commitments move apart from each other?

Note that answering such questions requires conceptual innovations. For instance, to be able to talk about convergence, we need a distance measure between epistemic states. Our model of RE suggests some ideas in this respect, but they need careful elaboration.

Some of the computer simulations run in the previous sub-projects produce results that can be re-used at this point. For instance, in sub-project 1, various trade-offs are studied for several sets of initial conditions. So, we can re-analyse and re-interpret the results and need not run new simulations. However, we expect that some possible sources of disagreement (e.g. iv) require additional simulations. These will first concentrate on the rules that we have specified in our present model of RE. Later, when sub-project 2 has worked out proposals for rules that can be followed by agents with bounded rationality, we'll consider such rules, too.

Box 3: Suggested simulation experiments – sub-project 3

Re-analyse the simulation results of sub-projects 1 and 2 by focusing on pairs of simulation runs that differ in terms of trade-offs, initial commitments etc. Investigate, in particular,

- whether agreement between global optima is typically larger than agreement between the corresponding initial commitments,
- how the evolution of mutual agreement – or disagreement? – in the course of equilibration processes looks like.

In this context, we can also address the worry that RE is too conservative. Conservativeness is intimately linked to consensus formation in the following way: If even agents who start with markedly different initial conditions converge on their views, RE is not conservative. We will take the opportunity to evaluate the results of our simulations to address the alleged conservativeness of RE too.

The results will provide us with a detailed picture of the power of RE to promote consensus formation. We will use the results to address the following issues:

1. Is the extent to which RE does or does not lead to consensus formation a problem for RE? Answering this question requires a careful analysis. We cannot expect that RE yields consensus under all circumstances. The question is rather under which conditions a lack of consensus is a pitfall of the method. Consider, for instance faults on the part of the agents (item iv. on the list). To some extent one may say that RE cannot be blamed for a disagreement that is caused by mistakes on the parts of the agents. But if minor mistakes make a huge difference, this strategy becomes implausible since some fault-tolerance is a reasonable requirement on methods.
2. What do our results imply for the current philosophical debate on disagreement? The epistemology of disagreement has recently attracted much attention (see e.g. Christensen & Lackey 2013, Cohnitz & Marques 2014 for recent collections and Frances & Matheson 2018 for an overview). So far, the debate has mainly focused on the consequences that we should draw from peer disagreement, e.g. epistemic modesty or scepticism, a main question being whether disagreement provides reasons to dispense judgment. But clearly, whether we should suspend judgement turns on the question of how the disagreement has arisen in the first place. We thus plan to use our results on disagreement and consensus formation to obtain a more nuanced view of the epistemic import of disagreement. Our working hypothesis is that suspension of judgement is less reasonable under circumstances for which RE likely leads to consensus.

Sub-project 4: Is it about real stuff? Matching metaphysics to RE

Suppose that RE is an appropriate method to obtain justified commitments in a given domain of discourse. What does this imply for our understanding of the domain? Can we give the domain a realist reading, according to which statements in the domain are literally true or false, and in part known to be true? Or does the use of RE favour a non-realist understanding, for instance a constructivist one?

The aim of this sub-project is to address these questions for the domain of discourse constituted by talk about normative practical reasons. We thus consider judgments of the type that Monica has a reason to invite her friends. Our guiding hypothesis is that any meta-ethical view about this discourse needs to offer an epistemology that explains how judgements about reasons can be known or at least be justified (see Peacocke 1999, Ch. 1 for such a requirement).

We plan to pursue a two-pronged strategy.

A) To understand whether RE can provide an exhaustive epistemology that fits a realist view of reasons, we'll first examine two influential *realist* accounts of reasons in which RE figures prominently. The accounts by M. Smith (1994) and T. M. Scanlon (2014) seem most interesting in this respect.[1]

In each case, we propose to examine the plausibility of the view by addressing the following two issues:

[1] The view of Smith has evolved since his (1994), see e.g. Smith (2007). Here, we concentrate on Smith (1994) because it prominently refers to RE.

- Each account is supposed to capture a common-sensical, pre-theoretical view about reasons. What consequences does this view imply for the epistemology, on the understanding of both authors? In particular, to what extent is consensus formation expected?
- Are these implications and expectations fulfilled as a matter of fact? Here we can use our model and in particular draw on results from sub-project 3.

Both accounts raise specific, but quite different issues:

Smith's use of RE is peculiar because he applies it particularly to desires (and not just to cognitive states). The idea is that "more rational selves" use the existing desires of real people and further develop them using RE to arrive at a desire what to do in the situation of the agent. Smith then proposes to use facts about what the advisors would recommend to analyse judgments about reasons. So far, Smith's position hasn't been analysed in depth with regard to his reference to RE. A closer analysis of Smith's view needs to answer the following questions:

- How can RE be applied to desires? In particular, what does it mean that desires are to be explained?
- How can the specific role of the advisor be captured using RE?

Scanlon's position is quite different. In contrast to Smith, he takes reasons to be unanalysable and suggests that facts about reasons constitute a distinct layer of reality. He prominently uses RE to argue for his position (Scanlon 2014, Ch. 4). In particular, he seems to think that RE provides an epistemology that explains we how can know about reasons (this at least is the way Altehenger et al. 2015 understand Scanlon). He also discusses the possibility that an application of RE offers an account of reasons that allows for the identification of an appropriate epistemology. This raises the following questions:

- What do unanalysability of reasons and the assumption of various layers of reality mean for the use of RE, in particular for the reference to background theories?
- How can a realist à la Scanlon explain that an epistemic agent knows a truth about reasons? In particular, does the explanation only appeal to RE or can facts about reasons play a significant role in the explanation (as some realists think they should)? If the latter is not the case, how can we explain the parallelism between the supposed truths about reasons and our alleged knowledge about them (see Altehenger et al. 2015 for a related criticism of Scanlon's)?

Altogether, we hope that a closer analysis of both positions allows for a more general assessment of the prospects for a realist interpretation of reasons that relies on RE.

It is possible to complement this analysis with yet further simulations of our model: In a dialectical structure, a consistent set of sentences from the pool can be assumed to be true, and we study to what extent full RE states and fixed points reached by following the rules at least approximate the truth (see Betz 2012 for this methodology). But such an analysis is not necessary if it turns out that consensus formation is already a problem for RE.

B) As a second part of our strategy, we plan to elaborate a version of *constructivism* that builds upon RE. Very roughly, a constructivist view about reasons holds that truth conditions of reason judgments can be spelled out by reference to a formal procedure or a specific point of view (Street 2010; see Rawls 1980 for a classical version of constructivism). As Street (2008:238) stresses, RE may play a crucial role for such a position if it is central for the procedure. RE would then be constitutive for the content of those judgments that constitute normative truths in the constructivist view. This avoids the criticism against Scanlon by Altehenger et al. (2015). We plan to follow this hint and to elaborate and discuss a version of constructivism in which RE is in fact constitutive in this sense. We hope in particular that such a position can use RE to make sense of the metaphor of construction.

It may first seem straight-forward to define a constructivist position that takes RE to be central. But there are challenges. For one thing, many constructivists, e.g. Rawls (1980), stick to some sort of objectivity (cf. Street 2012 and Way & Whiting 2017 for a different view). If the prospects of reaching consensus via RE can be shown to be dim, then the objectivist aspirations of some constructivists come under pressure because objectivity requires consensus formation at least under ideal circumstances. We can draw on our results from sub-project 3 to address this issue. It's interesting in this respect that the recent discussion about constructivism offers resources to evade the problem: One may argue that certain mental states (e.g. specific desires) are constitutive of agency and can thus be assumed for every agent (this claim is called constitutivism, see e.g. Smith 2013). We'll investigate how this move increases the chances of consensus formation under RE.

As another challenge, Schroeter (2004) has argued that a constructivist outlook and a commitment to RE exclude anti-theory, i.e., roughly the rejection of principles governing reasons. Consequently, constructivists need either reject anti-theory or Schroeter's analysis. We'll examine this challenge by closely looking at the theoretical virtues implicit in RE (cf. sub-project 1). At this point, it will be useful to consider Scanlon's view that RE cannot underwrite a constructivist view of moral reasons by giving a unifying account for them (Scanlon 2014, Ch. 4).

Box 4. Positions examined in sub-project 4 (rough characterizations):

- realism about reasons: Some of our judgements about reasons hold true independently of what we think, and are, or can be, known to be true.
- Meta-ethical constructivism: Whether or not a judgement about practical reasons is true depends on whether it can be constructed from a practical perspective or on the basis of an appropriate procedure.

Integration

To ultimately achieve the task of our research project, we'll carefully synthesize the results of the sub-projects. We'll put together what we've learned about the power of the method and its limitations. In particular, we'll check whether the free parameters in our model and the rules can be set such that

i. the trade-off of the various desiderata is plausible;
ii. the rules let non-ideal agents make significant progress in view of the desiderata;
iii. the results are stable;
iv. the process is fault-tolerant.

Only if we can identify a version of RE that fulfils all constraints, there is a version of RE that can be defended simultaneously against important challenges. We will use this version of RE to re-assess the most important objections against RE.

Interconnections between the sub-projects, methodology and collaboration

The sub-projects are closely related to each other, not only with respect to their topics, but also methodologically, because they involve investigations of the same model of RE and the application of similar techniques. The team members will thus be able to learn from each other and to join efforts in achieving their tasks. Furthermore, results of some sub-projects will be relevant for the other sub-projects. In particular, the results from sub-project 1 about plausible trade-offs can be used in the other sub-projects in order to focus on trade-offs that have been identified as plausible. The results of sub-project 2 about the RE process, in turn, will be relevant to sub-projects 3 (consensus formation) and 4 (meta-ethical consequences). Sub-project 4 will be informed by results from all other sub-projects. But there will also be an impact the other way around. For instance, results about the RE process (sub-project 2) and consensus formation (sub-project 3) may suggest a re-adjustment of the

desiderata in RE. To ensure that this potential for synergy is realized, it is important that the sub-projects are not carried out in isolation. To make sure that basic results from sub-project 1 are available early enough, sub-project 1 will start with the exploration of our model.

For the first three sub-projects, we plan to work with concrete examples of applications of RE, in particular from ethics. For instance, we wish to consider consensus formation for a specific ethical problem from consumer ethics. We'll also construct examples from the literature which illustrate paradigmatic differences between e.g. contractualist and utilitarian ethical theories. But we will not carry out another detailed case study, since this has already been at the centre of our previous research.

To foster the interactions between the Ph.D. students and to make sure that the results that are needed for other sub-projects are available in time, we'll define a couple of specific tasks that are supposed to be addressed by a team of 2 students. For instance, we plan that the students from sub-projects 2 and 3 join efforts to analyse and visualize the results from simulations of the RE process.

One – but not the only – crucial basis for the proposed research project is a formal model that we have developed for RE. The most important role of the model is to address expectations and fears about RE that could not yet be examined since RE wasn't properly elaborated. Despite this formal aspect, the project is primarily philosophical: it crucially relies on an informal clarification of RE, and its main goal is to scrutinize a philosophical method in order to address philosophical claims made about it.

The behaviour of the model can most easily be explored using computer simulations. We have finished a version of the simulation program, so no programming skills are expected from the Ph.D. students. What is rather crucial for success of the dissertation is a very good philosophical education. The analysis and the visualization of results that derive from many simulation runs needed for the project are not entirely straight-forward and require some new ideas on how to best organize the rich information implicit in the simulation outcomes. But this will not be a problem for students with a bit experience with e.g. data analysis. To make sure that students command the relevant skills, we will provide some specific training; in particular, we'll start the project with an intensive training week in Karlsruhe.

Since computer simulations start from very specific conditions, they need to be complemented with analytical methods (simple proofs) to obtain general results. We have already started research with analytic methods.

Although the proposed research relies to some part on the use of our model of RE, this model is not sacrosanct. We rather expect that we have to adapt and develop it further. One team member (see below for details) will be responsible to check how well the model works, to propose changes and to help implement them. Moreover, we are currently about to develop alternatives to the model. A master thesis project by Andreas Freivogel investigates the relationship to belief revision theory. We will take alternatives into account, if our original model doesn't live up to its aspirations. This flexibility is also crucial to ensure the Ph.D. enough autonomy in pursuing their projects.

Team members and their roles. We apply for funding of 4 Ph.D. students who work in the 4 sub-projects. The students are required to have a master in philosophy. For the dissertations 1 – 3, a minor in mathematics or a related field will be an advantage. Since we think that each Ph.D. student has to develop her/his own profile and to learn working as an autonomous researcher, we'll require them to put together an exposé for the individual dissertation.

So far, the software for the simulations of our model has been developed by GBe. An important aspect of his role will thus be to introduce the Ph.D.s to the computer code and to supervise them as far as the simulations are concerned. The Ph.D. students involved in subprojects 2 and 3 will be located in Karlsruhe with GBe as their first supervisor. In this way, the Ph.D. students will be able to profit from GBe's experience in designing simulations of argumentation under conditions of bounded

rationality and in investigating consensus formation (see Betz 2012). The other two Ph.D.s will be located at Berne and be supervised by CB, who commands special expertise in questions of epistemic values, meta-ethics and realism, relevant to subprojects 1 and 4 (see e.g. Beisbart 2007, 2008, under review). CB's main role will otherwise be to guide research with analytical methods (e.g. proofs about the existence of equilibria) and to ensure the connections with the philosophy of science and the meta-ethical literature. In addition, we apply for a further employee (50%, working at Berne), who will be responsible for the following tasks, which are essential to the project but cannot be assigned to the subprojects:

- make sure that the conceptual work and the simulations done in the sub-projects are properly related to the literature about RE;
- track challenges for the model and propose modifications, based on the philosophical debate about RE;
- synthesize and publish, together with the other project members, a survey of the results of the project as a contribution to the debate about RE;
- take an active part in the supervision of the Ph.D. students;
- coordinate project activities, especially the interaction between the students, organize workshops, and monitor the progress of the sub-projects;

Recruiting team members. The Ph.D. positions will be internationally advertised to make sure that we obtain the best possible candidates. Because many potential employers compete for Ph.D. students with a master in philosophy and some interest in formal methods or computer simulation, and because we seek to reduce the discrepancies in terms of Ph.D. salaries within our international project, we apply for full-time Ph.D. positions with DFG.

We are confident that we can fill all positions with competent students. CB is currently supervising the master theses by Andreas Freivogel and Sebastian Drosselmeier who work on topics closely related to RE. Both students are very good and excellent candidates for two Ph.D. positions in the project. Two other students have already indicated their interest to write a dissertation in the project.

The tasks assigned to the further employee can only be achieved by a person who has in-depth knowledge of both RE and our model, who is qualified to supervise Ph.D. students, and who possesses the management skills necessary for coordinating a multi-national project. We therefore propose to employ an experienced researcher, Georg Brun, with whom we have already collaborated. He has currently a permanent position at the University of Bern (Dozentur, 50%) and can invest 50% fte. in the project. GBr is an internationally renowned expert on RE, who has also twenty years of non-academic experience as a software engineer. He therefore has an ideal background that combines philosophy, computer technology and project management. His habilitation in philosophy (2015) qualifies him to supervise Ph.D. students.

Cooperation. To share the results from the sub-projects we are planning 2 project meetings per term that alternate between Karlsruhe (KA) and Berne (BE) and bring together all project members. The meetings will allow the students to present their results, occasionally feature talks by guests who work on closely related topics. The meetings will also be used to coordinate research activities.

We are planning that each Ph.D. student spends 1 term (about 4 months) and another 2 weeks at the other place. These "lab visits" are necessary for addressing tasks that are assigned to pairs of students and for close supervision e.g. about simulations (KA) or analytical results (BE). The research visits to KA are planned a bit earlier to make sure that the students finish their research with the simulations early enough.

There will be additional meetings of the local groups in BE and KA. During the first term of the project, we plan a weekly jour fixe at every place, where we discuss papers that help students to become familiar with the literature about RE. The jour fixe will also advertised as a colloquium for master students and hopefully attract additional students.

To coordinate the supervision of the Ph.D. students, each of the Karlsruhe students will have a second supervisor from Berne; likewise, at least one of the Berne students will be co-supervised by GBe (KA). GBr will co-supervise 2-3 Ph.D.s.

Software and outputs from simulation will be shared via a GIT platform hosted at KIT.

2.4 Schedule and milestones

Time	Work	Milestones
preparation phase before project start	Finish two paper drafts ("On the Plurality of Possible World-Views. RE and Pluralism" and "How Coherent Are Two Approaches Regarding Coherence with Each Other? RE and probabilistic coherence") presentations at GAP.10 advertisement of Ph.D. positions	
Months 1 – 3	Familiarization with RE and computer simulation program; training week in Karlsruhe; Ph.D. students complete their exposés	
Months 4 – 6	Analysis of the relevant literature and first practices with the computer program	
Months 7 – 12	First research for sub-projects, both BE students in KA	- 1st key publication about the weights of the desiderata of RE (sub-project 1)
Months 13 – 18	Research with focus on analysis of the model	- 1st international workshop on disagreement and philosophical method in Karlsruhe - 2nd key publication about consensus formation (sub-project 3)
Months 19 – 24	Research with a focus on conceptual issues, students from KA in BE	- 3rd key publication about equilibration under conditions of bounded rationality
Months 25 – 30	Research with focus on the interpretation of results	- 2nd international workshop on the power of RE in Bern - 4th key publication about meta-ethical consequences (sub-project 4)
Months 31 – 36	Finalization of dissertations	- dissertations - 5th key publication about consequences for our model of RE - 6th key publication about consequences for RE

We are planning 2 workshops:

- "More than agree to disagree? Disagreement and philosophical method" (KA): The aim of the workshop is to understand how philosophical methods may be used to resolve or at least understand disagreement. A call for papers welcomes contributions that may, or may not, have a formal aspect.
- "How powerful is reflective equilibrium? A critical assessment of a popular philosophical method" (BE): The aim of this workshop is to assemble defenders and critics of RE to understand the power and possible limitations of RE. Options for invited speakers: C. Z. Elgin (expert on RE), S. McGrath (critic of RE), J. Knobe or S. Nichols (defenders

of experimental philosophy), F. Tersman (proponent of RE), E. Olsson (expert on consensus formation and meta-ethical consequences).

2.5 Relevance and impact

Scientific relevance

This research project is located at the intersection between epistemology (formal and informal), ethics, meta-ethics and metaphilosophy. It thus is highly relevant for several philosophical discussions:

- We expect interesting results for the *epistemology of disagreement*. In particular, we use RE to obtain a systematic classification of possible reasons for disagreement. This will be particularly fruitful for the present discussion about the consequences of persistent disagreement.
- Our formal model and its possible modifications are likely to inspire new lines of research in *formal epistemology*. Note in particular that a lot of frameworks proposed in formal epistemology and related field concentrate on the question of how to handle empirical evidence. This is not typical for philosophy though, and our model will pioneer formal accounts of more philosophical methods.
- The proposed research is extremely important for the *methodology of ethics*. As indicated above, a lot of ethicists feel committed to RE as a method of ethical inquiry, although the method was so far barely used in a rigorous way. Our elaboration of the method, in particular for agents with bounded rationality, has the potential for a real methodological breakthrough in ethics and a lasting impact on work in all parts of ethics.
- Our research is also very important for *meta-philosophy* and the self-understanding of *philosophy*. Philosophy has always been a controversial discipline that is not so much shaped by consensus, but rather by disagreement. This research project will help to better understand the sources of disagreement and thus to arrive at a more reflective picture of philosophy.

As evident from our schedule, we plan 6 key publications, which will be submitted to internationally leading peer-reviewed journals such as *Mind, Philosophical Review, Erkenntnis, Synthese, Ethics, Ethical Theory and Moral Practice* or *the Journal of Ethics*. Apart from this, there will be 4 dissertations. Our Ph.D. students are expected to present their results at international conferences such as ECAP 2020 or GAP 2021. During the first year of the project, each Ph.D. student is expected to introduce her project at an international graduate conference.

Broader impact

Philosophical insight, in particular from more theoretical disciplines such as meta-ethics or epistemology, is not often easily applied in practice. Although quite theoretical, the proposed research goes amazingly far in the direction of applications. The reason is that our society faces huge ethical challenges (e.g. regarding climate change, migration, use of robots and artificial intelligence) that need profound moral deliberation because existing ethical concepts and theories can only give us limited advice in face of the challenges. The proposed research supports this type of deliberation by elaborating a method for addressing the challenges. Note in particular that we are interested to apply the method for realistic agents.

3. Bibliography

Altehenger, H., Gaus, S. & Menges, A. L. 2015, Being Realistic about Reflective Equilibrium, *Analysis* 75 (3):514–522.

Baumberger, C. forthcoming, Explicating Objectual Understanding: Taking Degrees Seriously. *Journal for General Philosophy of Science.* Available at https://www.academia.edu/34070486/Explicating_Objectual_Understanding_Taking_Degrees_Seriously.

Baumberger, C., Beisbart, C. & Brun, G. 2017, What is Understanding? An Overview of Recent Debates in Epistemology and Philosophy of Science, in: Grimm, S. R., Baumberger, C. & Ammon, S. (eds.), *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, New York, NY: Routledge-Taylor & Francis, 1–34.

Baumberger, C. & Brun, G. 2017, Dimensions of Objectual Understanding, in: Grimm, S. R., Baumberger, C. & Ammon, S. (eds.), *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, New York, NY: Routledge-Taylor & Francis, 165–89.

Beauchamp, T. L. & Childress, J. F. 2009, *Principles of Biomedical Ethics. 6th ed.* Oxford: Oxford University Press.

Beisbart, C. under review, Virtual Realism – Really Realism or Only Virtually So? Under review for *Disputatio.*

Beisbart, C. 2008, Praktische Gründe und moralische Prinzipien, in: Bohse, H. & Walter, S. (eds.), *Selected Papers Contributed to the Sections of GAP.6*, Paderborn: Mentis, 859–76.

Beisbart, C. 2007, *Handeln begründen. Motivation, Rationalität, Normativität*, Münster: Lit.

Beisbart, C., Betz, G. & Brun, G. under review, Spelling Out Reflective Equilibrium with a Simple Model, currently under review.

Betz, G. 2010, *Theorie dialektischer Strukturen,* Frankfurt a.M.: Klostermann.

Betz, G. 2012, *Debate Dynamics: How Controversy Improves Our Beliefs,* Synthese Library, Dordrecht: Springer 2012.

Bonevac, D. 2004, Reflection Without Equilibrium, *Journal of Philosophy* 101:363–88.

Brandt, R.B. 1985, The Concept of Rational Belief, *The Monist* 68:3–23.

Brink, D. O. 1989, *Moral Realism and the Foundations of Ethics,* Cambridge: Cambridge University Press.

Brink, D. 1987, Rawlsian Constructivism in Moral Theory, *Canadian Journal of Philosophy, 17*(1):71–90.

Brinkhuis, J. & V. Tikhomirov 2005, *Optimization: Insights and Applications*, Princeton/Oxford: Princeton University Press

Brun, G. 2009, Wer hat ein Problem mit irrationalen Präferenzen? Entscheidungstheorie und Überlegungsgleichgewicht, *Studia Philosophica* 68*:*11–41.

Brun, G.. 2012, Rival Logics, Disagreement and Reflective Equilibrium, in: Jäger, C. & Löffler, W. (eds), *Epistemology: Contexts, Values, Disagreement. Proceedings of the 34th International Ludwig Wittgenstein Symposium,* Frankfurt a.M.: Ontos, 355–68

Brun, G. 2014. Reflective Equilibrium Without Intuitions?, *Ethical Theory and Moral Practice*, 17:237–52.

Brun, G. 2017a, Conceptual Re-Engineering: From Explication to Reflective Equilibrium, *Synthese.* DOI 10.1007/s11229-017-1596-4

Brun, G. 2017b, Thought Experiments in Ethics, in: Stuart, M. T., Fehige, Y. & Brown, J. R. (eds), *The Routledge Companion to Thought Experiments,* Abingdon/New York: Routledge, 195–210.

Brun, G. 2018, Logical Expressivism, Logical Theory and the Critique of Inferences, *Synthese,* DOI 10.1007/s10670-015-9791–5.

Brun, G. & Betz, G. 2016, Analysing Practical Argumentation, in: Hansson, S. O. & Hirsch Hadorn, G. (eds), *The Argumentative Turn in Policy Analysis. Reasoning about Uncertainty,* Cham: Springer, 39–77.

van der Burg, W & van Willigenburg, T. (eds) 1998, *Reflective Equilibrium. Essays in Honour of Robert Heeger,* Dordrecht/Boston/London: Kluwer.

Cappelen, H. 2012, *Philosophy without Intuitions.* Oxford: Oxford University Press.

Cath, Y. 2016, Reflective Equilibrium, in: Cappelen, H., Gendler, T. & Hawthorne J. (eds.), *The Oxford Handbook of Philosophical Methodology,* Oxford University Press, 213–30.

Chalmers, D. J. 2015, Why Isn't There More Progress in Philosophy?, *Philosophy* 90 (1):3–31.

Christensen, D. & Lackey, J. (eds.) 2013, *The Epistemology of Disagreement: New Essays,* Oxford: Oxford University Press.

Cohnitz, D. & Marques, T. (eds.) 2014, Disagreements, *Erkenntnis* 79(S1):1-10.

Copp, D. 2012, Experiments, Intuitions and Methodology in Moral and Political Philosophy, in: Shafer-Landau, R. (ed.), *Oxford Studies in Metaethics.* Vol. 7, Oxford: Oxford University Press, 1–36.

Cummins, R. 1998, Reflection on Reflective Equilibrium, in: DePaul, M.R. & Ramsey, W. (eds.), *Rethinking Intuition.* Lanham: Rowman Littlefield, 113–27.

Daniels, N. 1979, Wide Reflective Equilibrium and Theory Acceptance in Ethics, *Journal of Philosophy* 76:256–282, reprinted in: Daniels 1996:21–46.

Daniels, N. 1980, On some Methods of Ethics and Linguistics, *Philosophical Studies* 37:21–36, reprinted in Daniels 1996:66–80.

Daniels, N. 1996, *Justice and Justification,* Cambridge: Cambridge University Press.

Daniels, N. 2018, Reflective Equilibrium, *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Zalta, E. N. (ed.), forthcoming URL = https://plato.stanford.edu/archives/spr2018/entries/reflective-equilibrium/.

de Maagt, S. 2017, Reflective Equilibrium and Moral Objectivity, *Inquiry : An Interdisciplinary Journal of Philosophy* 60(5):443–65.

DePaul, M. R. 2011, Methodological Issues. Reflective Equilibrium, in: Miller, C. (ed.), *The Continuum Companion to Ethics,* London: Continuum, lxxv–cv.

Doorn, N. 2009, Applying Rawlsian Approaches to Resolve Ethical Issues. Inventory and Setting of a Research Agenda, *Journal of Business Ethics* 91:127–43.

Elgin, C. Z. 1996, *Considered Judgment,* Princeton: Princeton University Press.

Elgin, C. Z. 2014, Non-Foundationalist Epistemology. Holism, Coherence, and Tenability, in: Steup, M., Turri, J & Sosa, E. (eds), *Contemporary Debates in Epistemology,* 2nd ed., Malden: Wiley, 244–55.

Feyerabend, P. 1975, *Against Method,* London: New Left Books.

Frances, B. & Matheson, J. 2018, Disagreement, *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Zalta, E. N. (ed.), URL = https://plato.stanford.edu/archives/spr2018/entries/disagreement/.

Gardiner, S. M. 2006. A Core Precautionary Principle, *The Journal of Political Philosophy* 14:33–60.

Gigerenzer, G. & Todd, P. M. 1999, *Simple Heuristics that Make Us Smart,* New York: Oxford University Press.

Goodman N. 1955/83, *Fact, Fiction, and Forecast,* 1st ed, Cambridge, MA: Harvard University Press, 4th ed. ibid.

Goodman, N. 1977, *The Structure of Appearance,* 3rd ed, Dordrecht: Reidel.

Hahn, S. 2000. *Überlegungsgleichgewicht(e). Prüfung einer Rechtfertigungsmetapher,* Freiburg/München: Alber.

Hahn, S. 2016. From Worked-out Practice to Justified Norms by Producing a Reflective Equilibrium, *Analyse & Kritik* 38:339–69.

Haslanger, S. 2012, *Resisting Reality. Social Construction and Social Critique,* Oxford: Oxford University Press.

Hempel, C. G. 1988, On the Cognitive Status and the Rationale of Scientific Methodology, in: Hempel, C. G., & Jeffrey, R. C. 2000, *Selected Philosophical Essays,* Cambridge: Cambridge University Press, 199–228.

Holmgren, M. 1987, Wide Refective Equilibrium and Objective Moral Truth, *Metaphilosophy* 18:108–24.

Huber, F. 2008, Assessing Theories, Bayes Style, *Synthese* 161:89–118.

Kappel, K. 2006, The Meta-Justification of Reflective Equilibrium, *Ethical Theory and Moral Practice* 9:131–47.

Kelly, T. & McGrath, S. 2010, Is Reflective Equilibrium Enough?, *Philosophical Perspectives* 24:325–59.

Knight, C. 2006, The Method of Reflective Equilibrium. Wide, Radical, Fallible, Plausible, *Philosophical Papers* 35:205–29.

Knobe, J. & Nichols, S. 2008, An Experimental Philosophy Manifesto, in: Knobe, J. & Nichols S. (eds.), *Experimental Philosophy*, Oxford University Press, 3–14.

Kuhn, T. S. 1962, *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

Kuhn, T. S. 1977, Objectivity, Value Judgment, and Theory Choice, in: Kuhn, T. S. 1977, *The Essential Tension.* Chicago: University of Chicago Press, 320–39.

Lacey, H. 1999, *Is Science Value Free? Values and Scientific Understanding*, London: Routledge.

Ladyman, J. 2002, *Understanding Philosophy of Science*, London: Routledge.

Lakatos, I. 1970, Falsification and the Methodology of Scientific Research Programmes, in: Lakatos, I. & Musgrave, A. (eds.), *Science and the Growth of Knowledge*, Cambridge: Cambridge University Press, 91–196.

Lange, K. 2004, *Optimization*, New York: Springer.

Lewis, D. 1994, Humean Supervenience Debugged, *Mind* 103(412):473–90.

Lewis, D. 1983, *Philosophical Papers.* Vol 1, New York: Oxford University Press.

Lewis, D. 1973, *Counterfactuals*, Blackwell: Oxford.

List, C. & Valentini, L. 2016, The Methodology of Political Theory, in: Cappelen, H., Gendler, T. S. & Hawthorne, J. (eds.), *The Oxford Handbook of Philosophical Methodology*, Oxford: Oxford University Press, 525–53.

Lycan, W. G. 2011, Epistemology and the Role of Intuitions, in: Bernecker, S. & D. Pritchard (eds.), *The Routledge Companion to Epistemology*, London: Routledge, 813–22.

Marino, P. 2010, Moral Rationalism and the Normative Status of Desiderative Coherence, *Journal of Moral Philosophy* 7(2):227–52.

McPherson, T. 2015, The Methodological Irrelevance of Reflective Equilibrium, in: C. Daley (ed.), *Palgrave Handbook of Philosophical Methodology*, 652–74.

Morreau, M. 2013, Mr. Fit, Mr. Simplicity and Mr. Scope: From Social Choice to Theory Choice, *Erkenntnis* (S6):1-16.

Nida-Rümelin, J. 1996, Theoretische und angewandte Ethik: Paradigmen, Begründungen, Bereiche, in: Nida- Rümelin, J. (ed.), *Angewandte Ethik. Ein Handbuch*, Stuttgart: Metzler, 2–85.

Okasha, S. 2011, Theory Choice and Social Choice: Kuhn versus Arrow, *Mind* 120(477):83–115.

Parfit, D. 2011, On What Matters, Oxford University Press: Oxford (Vols. 1 and 2).

Peacocke, C. 1999, *Being Known*, Clarendon Press: Oxford

Peregrin, J. & Svoboda, V. 2017, *Reflective Equilibrium and the Principles of Logical Analysis. Understanding the Laws of Logic*, New York: Routledge.

Rawls, J. 1971, *A Theory of Justice*, Cambridge, MA: Belknap Press.

Rawls, J. 1975, The Independence of Moral Theory, *Proceedings and Addresses of the American Philosophical Association* 48:5–22.

Rawls, J. 1980, Kantian Constructivism in Moral Theory: The Dewey Lectures 1980, *Journal of Philosophy* 77(9):515–72.

Resnik, M. D. 1997, *Mathematics as a Science of Patterns*, Oxford: Clarendon Press.

Sandin, P. 1999, Dimensions of the Precautionary Principle, *Human and Ecological Risk Assessment* 5:889–907.

Scanlon, P. 2014, *Being Realistic About Reasons*, Oxford: Oxford University Press.

Scanlon, T. M. 2003, Rawls on Justification, in: Freeman, S. (ed.). *The Cambridge Companion to Rawls.* Cambridge:

Cambridge University Press. 139–67.

Scanlon, T. M. 1998, *What We Owe to Each Other*, Harvard University Press: Cambridge (MA).

Scheffler, I. 1963, *The Anatomy of Inquiry.* New York: A. Knopf.

Schroeter, F. 2004, Reflective Equilibrium and Antitheory, *Nous*, *38*(1):110–34.

Simon, H. A. 2000, Bounded Rationality in Social Science: Today and Tomorrow, *Mind and Society* 1(1):25–39.

Singer, P. 2005, Ethics and Intuitions, *The Journal of Ethics* 9:331–52.

Singer, P. 1974, Sidgwick and Reflective Equilibrium, *The Monist* 58:490–517.

Slote, M. 2001, Moderation and Satisficing, in: Millgram, E. (ed.), *Varieties of Practical Reasoning*, Cambridge MA: MIT Press, 221–36.

Smith, M. 1994, *The Moral Problem*, Oxford: Blackwell.

Smith, M. 2007, In Defence of Ethics and the A Priori: Reply to Enoch, Heironymi, and Tannenbaum, *Philosophical Books* 48:136-149.

Smith, M. 2013, A Constitutivist Theory of Reasons: Its Promise and Parts, *Law, Ethics, and Philosophy* 1:9–30

Sosa, E. 2007, Experimental Philosophy and Philosophical Intuition, *Philosophical Studies* 132(1):99–107.

Spohn, W. 2002, The Many Facets of the Theory of Rationality, *Croatian Journal of Philosophy* 2:249–64.

Steel, D. 2015, *Philosophy and the Precautionary Principle*, Cambridge: Cambridge University Press.

Stegenga, J. 2013, An Impossibility Theorem for Amalgamating Evidence, *Synthese* 190 (12):2391–411.

Thagard, P. 1988, *Computational Philosophy of Science.* Cambridge, MA: MIT Press.

Stich, S.P. & R.E. Nisbett. 1980, Justification and the Psychology of Human Reasoning, *Philosophy of Science* 47:188–202

Street, S. 2008, Constructivism about Reasons, *Oxford Studies in Metaethics* 3:207–45.

Street, S. 2010, What is Constructivism in Ethics and Metaethics?, *Philosophy Compass* 5 (5):363–84.

Street S. 2012, Coming to Terms with Contingency: Humean Constructivism about Practical Reason. In: Lenman, J. & Shemmer, Y. (eds.), Constructivism in Practical Philosophy, Oxford University Press: Oxford, 40–59.

Strong, C. 2010, Theoretical and Practical Problems with Wide Reflective Equilibrium in Bioethics, *Theoretical Medicine and Bioethics* 31:123–40.

Swanton, C. 1992, *Freedom. A Coherence Theory,* Indianapolis: Hackett.

Tersman, F. 1993, *Reflective Equilibrium,* Stockholm: Almqvist and Wiksell.

Tersman, F. 2008, The Reliability of Moral Intuitions: A Challenge From Neuroscience, *Australasian Journal of Philosophy* 86:389–405.

Tersman, F. 2018, Recent Work on Reflective Equilibrium and Method in Ethics, *Philosophy Compass* 2018:e12493.

van Thiel, G. J. M. W. & van Delden, J. J. M. 2010, Reflective Equilibrium as a Normative Empirical Model, *Ethical Perspectives* 17:183–202.

von der Pfordten, D. (ed.) 2015, *Moralischer Realismus? Zur kohärentistischen Metaethik Julian Nida-Rümelins*, Münster: Mentis.

Walden, K. 2013, In Defense of Reflective Equilibrium, *Philosophical Studies 166:*243–56.

Way, J. & Whiting, D. 2017, Perspectivism and the Argument from Guidance, *Ethical Theory and Moral Practice* 20 (2):361–74.

Welch, J. R. 2014, *Moral Strata: Another Approach to Reflective Equilibrium,* Cham: Springer.

Williamson, T. 2007, *The Philosophy of Philosophy,* Malden: Blackwell.