Synthese (2024) 203:45
https://doi.org/10.1007/s11229-023-04466-y

ORIGINAL RESEARCH

Probabilifying reflective equilibrium

Finnur Dellsén[1,2,3]

Received: 25 March 2023 / Accepted: 16 December 2023 / Published online: 25 January 2024

Abstract
This paper aims to flesh out the celebrated notion of *reflective equilibrium* within a probabilistic framework for epistemic rationality. On the account developed here, an agent's attitudes are in reflective equilibrium when there is a certain sort of harmony between the agent's *credences*, on the one hand, and what the agent *accepts*, on the other hand. Somewhat more precisely, reflective equilibrium is taken to consist in the agent accepting, or being prepared to accept, all and only claims that follow from a maximally comprehensive theory that is more probable than any other such theory. Drawing on previous work, the paper shows that when an agent is in reflective equilibrium in this sense, the set of claims they accept or are prepared to accept is bound to be logically consistent and closed under logical implication. The paper also argues that this account can explain various features of philosophical argumentation in which the notion of reflective equilibrium features centrally, such as the emphasis on evaluating philosophical theories holistically rather than in a piecemeal fashion.

Keywords Reflective equilibrium · Probability · Optimality model · Acceptance · Credence

1 Introduction

Following Goodman (1955) and Rawls (1999), it is often suggested that many philosophical argumentats proceed, or should proceed, by employing the *method of reflective equilibrium*. To a first approximation, this method involves reflecting on and revising a set of particular and general judgments, which may initially conflict with

✉ Finnur Dellsén
fud@hi.is

1 Faculty of Philosophy, History and Archaeology, University of Iceland, 101 Reykjavik, Iceland

2 Department of Philosophy, Law, and International Studies, Inland Norway University of Applied Sciences, 2624 Lillehammer, Norway

3 Department of Philosophy, Classics, History and of Art and Ideas, University of Oslo, 0315 Oslo, Norway

each other, until they are eventually brought into a state in which such conflicts have been resolved, i.e. reflective equilibrium. For example, Goodman famously argued that the only justification possible for a system of rules for inductive or deductive inference lies in "the agreement achieved [by] making mutual adjustments between rules and accepted inferences" (Goodman, 1955, p. 64). Similarly, Rawls suggests that a person who adopts a conception justice such as his own should have "weighed various proposed conceptions and [...] either revised his judgments to accord with one of them or held fast to his initial convictions" (Rawls, 1999, p. 43).

For our purposes, it is important to distinguish between the *method* of reflective equilibrium, and its desired end-point, the *state* of reflective equilibrium. The former is a method for reaching the latter, so the two are closely related. However, to formulate and defend an account of the method of reflective equilibrium is in some ways a much more ambitious task than to formulate and defend a an account of the state of reflective equilibrium. After all, the former would presumably involve a general description not only of the desired end-point of the process recommended by the method, but also some prescription for how to carry out the step-by-step process of continually revising one's judgments in light of one another until that desired end-point is reached. In this paper, I will not be directly concerned with this more ambitious project in so far as it purports to tell us how to carry out this step-by-step process.

Rather, I focus here on the more modest task of formulating and defending an account of the state of reflective equilibrium, i.e. the desired end-point of the method.[1] The account I propose is in some ways speculative, and no doubt wrongheaded in some of its details, but I hope that the core idea is a worthwhile contribution to current thinking about (the state of) reflective equilibrium.[2] That core idea is, roughly, that reflective equilibrium consists in a certain type of harmony between what an agent *accepts*, on the one hand, and the agent's *credences*, on the other hand—where acceptance is a binary attitude that roughly consists in having a policy of including a proposition among one's premises in a particular context. Specifically, I'll suggest that reflective equilibrium can be explicated as an agent accepting, or being prepared to accept, precisely those proposition that follow from a maximally comprehensive theory that is, by the lights of the agent's credences, more probable than any alternative such theory. In this way, I am proposing a 'probabilification' of reflective equilibrium of a sort that I have not seen elsewhere.[3]

The paper proceeds as follows. I start by laying out several good-making features of an account of reflective equilibrium (Sect. 2). I then turn to discussing two notions that play starring roles in the idea of reflective equilibrium, namely what I shall refer to as 'judgments' and 'theories' (Sect. 3). This serves as a setup for a careful formulation

[1] To be sure, this will involve making certain (hopefully well-motivated) assumptions about the starting-point of reflective equilibrium as well, but it will not involve any prescription regarding the step-by-step process of getting from the starting-point to the end-point.

[2] From now on, unless otherwise indicated explicitly, 'reflective equilibrium' refers exclusively to the *state* (rather than the *method*) of reflective equilibrium.

[3] The only other discussion of reflective equilibrium from a probabilistic perspective of which I am aware is contained in an unpublished Ph.D.-thesis by Vallinder (2018). Although Vallinder's approach is very interesting, it is also quite different from the one taken here, so further discussion of it here would take us too far afield.

of a probabilistic model of reflective equilibrium, which I call the *Optimality Model of Reflective Equilibrium* (Sect. 4). Having laid out that model, I return to desiderata outlined previously and explain how the model satisfies these desiderata (Sect. 5). I conclude by briefly discussing some of the model's upshots and limitations (Sect. 6).

2 What do we want from reflective equilibrium?

The notion of reflective equilibrium as developed by its most influential proponents, such as Goodman and Rawls—and, more recently, Daniels (1979) and Tersman (1993)—remains open-ended along several dimensions. Moreover, we should be prepared to consider the possibility that some of the features attributed to reflective equilibrium by its most influential proponents are peripheral or even dispensable. Thus, instead of systematically reviewing the various claims about, or accounts of, reflective equilibrium that are present in the extant literature, I will rather attempt to lay down some good-making features, or desiderata, that I hope we can all agree that an account of reflective equilibrium should ideally satisfy. By referring to these good-making features as 'desiderata', I do not mean to suggest that an account that fails to satisfy them would be a non-starter—after all, we cannot rule out beforehand that they cannot all be simultaneously satisfied—only that they should serve as a list of desirable features that can serve as our guide when developing and evaluating a probabilistic account of reflective equilibrium in later sections.

2.1 Making sense of philosophical argumentation

The first and most important such desideratum is that our account of reflective equilibrium should help to make sense of philosophical argumentation, at least in those cases in which philosophers explicitly appeal to 'reflective equilibrium'. More precisely, the account should be applicable to as many cases as possible in which philosophers claim to have reached a state of reflective equilibrium, or in which they claim to use the method of reflective equilibrium. Of course, an account need not imply that every instance in which philosophers appeal to reflective equilibrium is a legitimate instance thereof, since philosophers may themselves be confused about their own methodological practice to some extent. However, certain core features of philosophical arguments that appeal to reflective equilibrium should arguably be present in any adequate account thereof.

In particular, since much of philosophical reasoning appears to involve appealing to pre-theoretic judgments, i.e. what some would call 'intuitions', this practice is something that an account of reflective equilibrium should at least allow for and preferably help make sense of. For example, such an account should help explain why it is that debates about utilitarianism appeal to people's pre-theoretic judgments about hypothetical trolley cases, in which we are presented with the option of pulling a lever to divert a trolley away from a track on which there are five people and onto a track on which there is one person (Rechnitzer, 2022b). How are our pre-theoretic judgments about such cases relevant for whether we should end up accepting utilitarianism? After

all, such judgments are just opinions. Relatedly, it seems that pre-theoretic judgments can sometimes legitimately be 'overridden' by an overarching theory with which they conflict, as Goodman and Rawls both suggest. But how is this possible if the overarching theories are themselves justified in virtue of fitting those very same pre-theoretic judgments? An account of reflective equilibrium should help us answer these sorts of questions.

Another feature of philosophical argumentation that an account of reflective equilibrium ought to be consistent with, and ideally help to explain, is the appeal to explanatory virtues in the evaluation of philosophical theories. In particular, philosophers frequently assume that, all other things being equal, a simpler theory should be preferred to a more complex one, a unifying theory should be preferred to a collection of separate theories, and so on for various other explanatory virtues. To be sure, this preference for explanatorily virtuous theories is not a special feature of reflective equilibrium but rather a general feature of philosophical argumentation (assuming for the sake of the argument that not all philosophical argumentation employs the method of reflective equilibrium).[4] However, it would surely be unfortunate if our account of reflective equilibrium left unexplained, or even conflicted with, the common preference among philosophers for explanatorily virtuous theories. For example, if all else is equal between two philosophical theories T_1 and T_2 except that T_2 posits two distinct types of fundamental substances (e.g. minds and matter) where T_1 posits only one (e.g. only matter), then our account of reflective equilibrium should at least be consistent with, and ideally help to explain, our preference for T_1 over T_2.

2.2 Fit with epistemological framework(s)

A second desideratum is that an account of reflective equilibrium should allow us either to situate reflective equilibrium within an epistemological framework (or collection of such frameworks) that we have independent reasons to endorse, or at least to see how reflective equilibrium might complement it (or them) in some way. The epistemological framework(s) in question should ideally be as informative as possible, for example in not just claiming that beliefs are justified in virtue of 'cohering' with other beliefs, but also in specifying what exactly is involved in *cohering* with other beliefs, what sort of *beliefs* are at issue, and so forth. In short, reflective equilibrium ought to fit with as much as possible of the epistemological framework(s) one accepts for other purposes, e.g. in accounting for what's rational to believe in scientific or everyday contexts, rather than standing alone as some sort of ad hoc epistemological exception which applies only to philosophical arguments or some subset thereof.

Why take this to be a desideratum for an account of reflective equilibrium? What's wrong with an 'exceptionalist' account of reflective equilibrium on which it comes with its own special epistemology that doesn't fit inside, or even complement, one's other epistemological framework(s)? Two reasons seem to me to be most decisive. First, such an exceptionalist account would arguably be somewhat self-undermining, in that it seems impossible for one's attitudes towards reflective equilibrium itself

[4] With that said, it is not uncommon to associate reflective equilibrium especially with explanatory coherence and/or a preference for explanatorily virtuous theories (e.g. Føllesdal, 2005; Lycan, 1985).

to be in reflective equilibrium with the epistemological framework(s) one accepts in other contexts if the former cannot be situated within, or be shown to complement, the latter. Put differently, an exceptionalist account of reflective equilibrium would imply a sort of disconnect between one's account of reflective equilibrium and one's general epistemological framework(s) that reflective equilibrium itself would require one to resolve. The resolution of such a conflict would motivate a revision of either the account of reflective equilibrium or the general epistemological framework(s)—or both; either way, the former would indeed be situated within, or complement, the latter.

A second reason why it is desirable to be able to fit reflective equilibrium within more general epistemological framework(s) concerns worries about whether achieving reflective equilibrium really is as significant as its advocates take it to be. This is not to deny that numerous philosophers, including such titans of analytic philosophy as Goodman and Rawls, have made reflective equilibrium into a cornerstone of their philosophical methodology. That much is undeniable. What is not undeniable, indeed frequently denied, is that reflective equilibrium *deserves* to play such a central role in philosophical methodology (e.g., Kelly & McGrath, 2010; McPherson, 2015; Singer, 1974). The concern, in other words, is that reflective equilibrium may only be a description of what we philosophers in fact often do when we argue for philosophical theories, and not also a plausible normative account of how we *should* be arguing for such theories. If it's only the former, then there is a real danger that the widespread endorsement of reflective equilibrium in analytic philosophy merely serves to shroud us from a methodological rot at the heart of much philosophical theorizing.

2.3 Deductive cogency of resulting theories

A third desideratum concerns what logical features a set of theories must have if they are to qualify as being in reflective equilibrium. There are two related, but strictly speaking separable, logical features that I suggest should characterize such theories. The first and most straightforward is that the theories should be *deductively consistent*, both individually and with each other. To illustrate, this entails that if orthodox utilitarianism is inconsistent with positing a morally significant distinction between killing someone and letting that person die, in the sense that this could by itself make the difference between morally right and wrong action, then our account of reflective equilibrium should entail that someone who accepts both orthodox utilitarianism and the moral significance of killing versus letting-die is not (yet) in a state of reflective equilibrium; rather, they must revise one or both of these ideas in order to achieve reflective equilibrium.

A second logical feature that I suggest characterizes a set of theories in reflective equilibrium is *deductive closability*. A set of theories is deductively closable for some person *S* just in case *S*—either explicitly or, more commonly, implicitly—is prepared to accept what follows deductively from the theories she accepts.[5] It is hard to make sense of philosophical argumentation if we do not impose a closability requirement on

[5] Here I am using 'prepared to accept' in a weak sense on which one can be prepared to accept something that one doesn't—or couldn't, given one's human limitations—fully understand, e.g. a very complex logical theorem. Specifically, I shall assume that being 'prepared to accept' something implies that if one *were* to fully understand and consider it, one *would* accept it. Nevertheless, the requirement of deductive closability

the theories we accept. After all, a common form of objection to a philosophical theory is that is has some undesirable implications, e.g. that it conflicts with an intuitive verdict in some hypothetical example. If theories in reflective equilibrium needn't satisfy deductive closability, then there is nothing to prevent a philosopher from claiming that their theories are in reflective equilibrium even though they refuse to accept an obvious deductive implication of those theories. For example, such a philosopher could accept a justified-true-belief theory of knowledge and yet refuse to accept that the subjects in Gettier's counterexamples have knowledge (without even denying that the former implies the latter!).

These two requirements—of deductive consistency and closability—are sometimes jointly referred to as 'deductive cogency' in other contexts (e.g., Christensen, 2004, Chap. 3; Dellsén, 2018; Kaplan, 1996, Chap. 3). So we can say here that reflective equilibrium demands that the resulting theories be *deductively cogent*. To be clear, this is not to say that the method of reflective equilibrium demands deductive cogency at every stage of the process. In particular, I see no reason to demand deductive cogency of the inputs and working posits of the method of reflective equilibrium, i.e. what Rawls (1999, p. 42) refers to as (considered) *judgments*; only that this demand applies to the resulting *theories* that we end up accepting (see Sect. 3 for more on this distinction). Indeed, transforming a set of claims that are initially inconsistent or unclosable into a set that is deductively cogent seems to be precisely the sort of process that proponents of reflective equilibrium are envisioning. For this reason, we should at least allow for our initial judgments to fail to be consistent or closable even when the theories we end up accepting are deductively cogent.

2.4 Summary: three desiderata

In sum, then, it seems that an account of reflective equilibrium should ideally have the following three features. First, it should make sense of the ways in which philosophers who appeal to reflective equilibrium argue for their theories, perhaps especially when they appeal to pre-theoretic judgments (i.e., 'intuitions') that they are nevertheless prepared to revise later on. Second, an account of reflective equilibrium should ideally cohere with or complement—and thus be in reflective equilibrium with—epistemological frameworks that are independently plausible by our lights. Third, the account must explicitly or implicitly require that the set of theories accepted by an agent in reflective equilibrium is deductively consistent and such that the agent would not refuse to accept anything that deductively follows from it. These three requirements seem rather innocuous at first blush; however, as we shall see below (Sect. 5),

might seem overly demanding in that, since any theory will have an infinite number of deductive implications, it implies that agents should be prepared to accept infinitely many claims. However, note that the requirement of deductive closability merely requires that *S* be *prepared* to accept anything that follows deductively from something she already accepts, and there is nothing paradoxical (or even unusual) about being prepared to do an infinite number of different things. Moreover, if *S* discovers that she is not prepared to accept some particular implication of claims theories she already accepts, then she may respond by discarding one of the theories rather than slavishly accepting its implications. This is of course what happens when we realize—or are made to realize—that some combinations of our views lead to an absurd conclusion in an *reductio ad absurdum*.

they suggest that our account of reflective equilibrium should have a rather specific structure.

3 From judgments to theories

In this section, I consider what we may roughly think of as the 'inputs' and 'outputs' of the method of reflective equilibrium, i.e., the judgments (commitments, convictions) that one starts out with and then revises, on the one hand, and the resulting theories (principles, accounts) that one ends up accepting, on the other hand. My focus will be on the little-discussed question of what sort of propositional attitude realizes these functional roles. Regarding the latter, I will draw on Elgin's (2017, Chaps. 2–3) suggestion that accepted theories should be associated with what what one takes for granted in the context of understanding something. Regarding the former, I will suggest that such judgments are simply the agent's credences.[6]

3.1 Judgments as credences

As Kauppinen and Hirvelä (2023, p. 1) nicely put it in a recent survey article on reflective equilibrium, "[n]o one ever begins ethical inquiry without already have many ethical convictions." These initial 'convictions', which are more commonly referred to as 'judgments' (or 'commitments') by other authors, form the starting points of the method of reflective equilibrium. As the method proceeds, these judgments are typically revised or developed in light of each other, in a sort of step-by-step process in which one works back and forth between them. That's the functional role of 'judgments' in the method of reflective equilibrium. But what type of propositional attitude, exactly, should be taken to realize this functional role? What are these 'judgments', really?

My proposal is simple. The attitudes that form the inputs and working posits of the method of reflective equilibrium in this way, i.e. the agent's 'judgments', can be identified with their *credences*. A credence in a proposition is a type of belief in the proposition that comes in some specific degree of strength. Such degrees of strength may range from a maximum value of 1 (representing absolute certainty that the proposition is true) to a minimum value of 0 (representing absolute certainty that the proposition is false). Credences may be *precise*, in which case their value is a single real number within in this range, $c \in [0, 1]$; or *imprecise*, in which case they are an interval within the range, $[c_1, c_2] \subseteq [0, 1]$. What I am suggesting, then, is

[6] As this indicates, I will assume throughout that the attitudes that come to be in reflective equilibrium are directed towards *propositions*, which are by definition truth-apt, i.e. capable of being true and false—albeit perhaps only in a deflationary sense of these terms. One might object to this assumption that reflective equilibrium should also apply to various purely motivational and non-propositional mental states, such as one's values and goals. In response, let me first note that although a value or goal is indeed not propositional (and thus not truth-apt) in and of itself, there are various propositions *about* these values and goals to which reflective equilibrium might easily apply, e.g. of the form 'Goal G is worth having' or 'Value V_1 is more important than value V_2'. I take such propositions to be perfectly truth-apt—although, again, perhaps only in a deflationary sense of 'true'. This may commit me to a rejecting some forms of non-cognitivism about normative sentences, but that's a commitment I am happy to take on.

that the method of reflective equilibrium takes as its input not simply a collection of propositions that the agent outright believes, but also the degree to which they believe each proposition, i.e. their (precise or imprecise) credences in these propositions.

There are two main reasons why I suggest we identify 'judgments' with credences in this way. The first is that it seems undeniable that both actual and ideal agents do have credences in one form or another. Indeed, there is an unusually widespread consensus in recent epistemology—unusual for philosophy, that is—that positing credences is necessary in order to describe both actual and ideal agents from an epistemic point of view. After all, of the many things that we have beliefs about, we certainly seem to believe some things more strongly than others, and some things a lot more strongly than others (and so forth). Furthermore, in light of the fact that our evidence can clearly provide more and less support for different propositions, it seems that an epistemically ideal agent would proportion their beliefs to the evidence accordingly, by having different degrees of belief, i.e. credences, in those propositions. So if there is any type of propositional attitude that must be taken to exist, both in actual and ideal agents, credence is arguably the best candidate available. Since the method of reflective equilibrium can only take as its inputs propositional attitudes that actually exist in us, credences seem to be a particularly good candidate for serving that function.

A second reason to identify 'judgments' with credences is that the former, like the latter, are clearly a matter of degree. As Brun (2014, p. 240) notes, "[j]udgements [...] include a propositional attitude involving a certain degree of commitment, which need not be definite or unwavering, but can also be minimal or feeble." Even Rawls (1999, p. 42) mentions that we should begin the process of reflective equilibrium by discarding "those judgments made with hesitation, or in which we have little confidence."[7] These comments presuppose that 'judgments' are a matter of degree, much like credences are defined to be (and in contrast to various other propositional attitudes, e.g. outright beliefs and acceptances). And rightly so, I would add, because some of the inputs into reflective equilibrium are surely directed towards propositions in which we have a great deal of confidence (e.g., that it's morally wrong to kill an innocent person), while we are less confident in others (e.g., that it's morally wrong to kill a mass murderer)—even when we are in some sense committed to both (albeit to different degrees).

Indeed, a model of reflective equilibrium that failed to respect such differences in the degree to which we are committed to different claims would arguably be unable to account for quite basic features of how philosophical argumentation actually proceeds (and, I would add, ought to proceed). After all, it seems undeniable that our degree of commitment to a particular 'judgment' should influence the extent to which revising it counts against the theory we end up accepting. For example, suppose we are faced with a choice between two ethical theories that are alike in all relevant respects except that the first (but not the second) implies that it's morally wrong to kill an innocent person while the second (but not the first) implies that it's morally wrong to kill a mass murderer. In this example, I take it that even those of us who would in some sense be

[7] Similarly, Goodman (1952) and Scheffler (1954) both endorse the idea that the initial commitments from which other attitudes are justified often fall short of being certain or definite. To be clear, both papers were written before the label 'reflective equilibrium' had been coined, but both are arguably describing early versions of the method of reflective equilibrium.

inclined to 'judge' both implications as correct would still prefer the first theory to the second. We are, so to speak, *more* committed to the first implication than to the second. All of this makes complete sense if we identify 'judgments' with an agent's credences, which are—by definition—a matter of degree.

I wish to emphasize that although the method of reflective equilibrium begins with a set of initial judgments—which I'm identifying with the agent's credences in the relevant propositions—these judgments can, and typically will be, *revised* as the method proceeds. Proponents of reflective equilibrium, such as Daniels (1979, §1), see this process as a kind of "working back and forth among our considered judgments [...] revising any of these elements wherever necessary". As I have noted, I will not have much to say about this process, focusing instead on the resulting state of reflective equilibrium. However, as I'll explain below (see Sect. 4), my model assumes that an agent's judgments in the state of reflective equilibrium, i.e. her relevant credences, have at that point been made probabilistically coherent. Since I am not here offering an account of the process involved in the method of reflective equilibrium, I will not put forward a proposal for *how* an agent could or should make their credences probabilistically coherent as part of this process. With that said, I am hopeful that recent work on degrees of probabilistic coherence (De Bona & Staffel, 2018; Staffel, 2019, esp. pp. 96–151) already contains the seeds of plausible accounts of this aspect of reflective equilibrium as I am conceiving of it.

3.2 Theories as acceptances

Let us next consider the outcomes of the method of reflective equilibrium, i.e. what many authors refer to as 'theories' (or 'principles', 'accounts'). These are, of course, the ideas for which philosophers are arguing in appealing to reflective equilibrium, such as Rawls's theory of justice or Goodman's theory of projectable predicates. These theories are often ambitious claims about general phenomena, such as what constitutes a just society or which predicates support inductive inferences. However, it's often noted that 'theories' cannot be distinguished from 'judgments' by their generality. After all, some of the 'judgments' that are formed prior to the process of reflective equilibrium, and which serve as inputs into that process, are perfectly general (e.g. that it's always wrong to kill other people). Furthermore, it's also possible for philosophical 'theories' to concern quite specific claims or phenomena (e.g. that God exists). So what, then, characterizes the 'theories' that form the output of reflective equilibrium, and how are they distinct from the 'judgments' that serve as the input and working posits of the method?

On this issue I take my cue from Catherine Elgin's work on the relationship between reflective equilibrium, understanding, and acceptance (see esp. Elgin, 2017, Chaps. 2–4, but also Elgin, 1996, Chap. 4; 2006; 2007; 2009), which has been influential in the recent literature (see, e.g., Baumberger & Brun, 2021; Jäger & Malfatti, 2021; Kauppinen & Hirvelä, 2023). In brief, Elgin and others suggest that reflective equilibrium is a method for gaining *understanding*—rather than, say, *knowledge*—of the objects

or phenomena about which you have judgments and form theories.[8] One important motivation for this view comes from noticing that understanding something seems to require a holistic representation of it and how it 'hangs together' with other things, e.g. by grasping how it depends on other things and how other things depend on it (Dellsén, 2020; Grimm, 2014; Kim, 1974; Kvanvig, 2003). This type of holistic representation also seems to be the goal of the process of reflective equilibrium, since it does not merely involve having correct representations of various isolated facts, but rather a systematic and unified representation of all of the relevant facts and how they relate to each other.

Now, if the goal of reflective equilibrium is understanding, then what does this imply about the 'theories' that serve as the outputs of the method? Well, that depends on what sort of propositional attitude is involved in understanding. Here again I take my cue from Elgin (2004, 2017)—as well as Baumberger (2018) and my earlier self (Dellsén 2017, 2018)—who suggest that the propositional attitude involved in understanding is a kind of *acceptance* in Cohen's (1992) sense of the term. According to Cohen's definition, to accept that P is "to have or adopt a policy of deeming, positing, or postulating that [P] – i.e. of including [P] among one's premises for deciding what to do or think in a particular context, whether or not one feels it to be true" (Cohen, 1992, p. 4; cf. Elgin, 2017, p. 19). Since the 'context' in which we are interested here is that of understanding something, we can say that the kind of acceptance of interest to us—which I have previously called *noetic acceptance* (Dellsén, 2018, p. 3134)—consists in having a policy of including P among one's premises in the context of understanding something. Importantly for our purposes, noetically accepting P does not by itself require or imply that one believes that P, or indeed that one's credence in P exceeds any specific threshold (Dellsén, 2018, pp. 3131–3132; see also Cohen, 1992, pp. 108–116).[9]

My suggestion, then, is that the type of propositional attitude that serves as the outcome of the process of reflective equilibrium is acceptance—or, more precisely, noetic acceptance. This attitude consist in having a policy of including the relevant proposition among one's premises for deciding what to do or think in the context of

[8] It is widely acknowledged that understanding and knowledge are different epistemic states, although it's controversial to what extent they may be related. Some have argued that understanding is a *species* of propositional knowledge (Grimm, 2006; Kelp, 2017; Khalifa, 2017; Sliwa, 2015), while several others explicitly reject a knowledge-based analysis of understanding—by arguing, for example, that understanding need not involve epistemic justification of the traditional sort (Dellsén, 2017; Hills, 2016), that it is immune to Gettierization (Kvanvig, 2003; Pritchard, 2009), or that it can essentially involve idealizations, which are false, in a way that knowledge cannot (Elgin, 2007; Mizrahi, 2012). If this latter set of views is on the right track, then Elgin's suggestion of viewing reflective equilibrium as a method for gaining understanding would nicely explain why it would be necessary to supplement the various established methods for obtaining knowledge, such as deductive and abductive argumentation, with the comparatively rather speculative method of reflective equilibrium. The latter would be a method specifically designed for obtaining understanding, rather than knowledge, and since the two states are distinct we shouldn't expect the methods for achieving these states to be identical.

[9] Of course, to say that understanding involves acceptance rather than belief (or credence) is not to say that belief (credence) could not play any role at all in how or why an agent comes to understand something. Rather, it means that belief (credence) is not *necessary* for understanding—that one *could* understand something without believing (having any specific credence in) the propositions on which the understanding is based.

understanding something. With that said, for the purposes of this paper, it would make little difference if one jettisoned the idea that the method of reflective equilibrium aims to produce understanding specifically (rather than, say, knowledge)—as long as one goes along with the (distinct, but to my mind related) idea that the output of the method is a kind of acceptance in Cohen's sense. For example, one could have a view of the method of reflective equilibrium on which it simply aims at true theories, but take acceptance to be the type of attitude one has towards such theories. Indeed, one could even suppose that reflective equilibrium aims at a type of knowledge, but argue that such knowledge involves acceptance rather than belief—as Cohen (1992, pp. 86–100) himself suggested is true of *scientific* knowledge. For this reason, I will not assume in what follows that the type of acceptance that serves as the output of reflective equilibrium must be *noetic* acceptance.

At any rate, one important feature of acceptance, in Cohen's sense of the term, is that it is a *binary* (on-or-off) type of propositional attitude. One either accepts P, or one does not, because P either is included among one's premises in some context, or it isn't. Thus, on the view I am sketching here, a given proposition P either is included among the 'theories' that one ends up with in reflective equilibrium, or it isn't. This, I submit, is a *feature* of the view rather than a *bug*, for two main reasons.

First, empirical research in psychology suggests that humans in general possess both degreed and binary propositional attitudes, neither of which can be reduced to the other without remainder (see Weisberg, 2020, and references therein). Although these results are tentative, because the empirical work is still nascent, it seems clear that a binary attitude very much like acceptance—involving, among other things, a commitment to use the accepted proposition as a premise in reasoning—is part of our human psychology, whether we like it or not. Furthermore, and perhaps more to the point, it seems clear that the philosophers who appeal to reflective equilibrium do in fact—at least normally—endorse the resulting theories in a similarly binary (on-or-off) sort of way. For example, Rawls (1999) did not merely recommend assigning a high degree of credence or plausibility to his theory of justice as fairness, but rather argued for some form of binary endorsement of the view. Although we can easily conceive of a method that ends up delivering degreed rather than binary conclusions, it's fair to say that this would be a novelty in philosophical argumentation and certainly not what reflective equilibrium is ordinary (if indeed ever) taken to involve.[10]

Second, I agree with a number of authors who have argued that positing a binary type of belief-like attitude, such as Cohen's acceptance, is necessary to make normative sense of various epistemic practices in which we regularly engage, especially as philosophers (see, e.g., Kaplan, 1996, 2013; Roorda, 1997; van Fraassen, 1995; see also Dellsén, 2018). In particular, it seems that at least in certain contexts, e.g. in the philosophy seminar room, we demand—of ourselves and others—that the attitudes one endorses be logically consistent; and, moreover, that one be prepared to endorse the logical consequences of what one already endorses (or else take back the initial

[10] As I explain below (see Sect. 4), having this type of binary attitude towards a philosophical theory is perfectly compatible with having a less than maximal—indeed, possibly a very low—credence in that theory. For this reason, it is also perfectly natural to express doubts about the philosophical theories one accepts. So to say that acceptance is binary is by no means to say that it involves some sort of dogmatic commitment to the accepted theory.

endorsement). In short, we demand deductive cogency in the sense described above (see Sect. 2.3). However, it would make little sense to demand deductive cogency of agents in reflective equilibrium if they did not even possess binary belief-like attitudes, e.g. of the sort Cohen calls acceptance. After all, deductive cogency is not a requirement on graded attitudes like crendences, nor is it a plausible demand on credences of a particular strength, as the lottery and preface paradoxes show (Kyburg, 1961; Makinson, 1965; for discussion, see Christensen, 2004).[11]

3.3 Summary: judging versus accepting

Let us take stock. I have suggested that the inputs and working posits of the method of reflective equilibrium, i.e. the 'judgments' with which one starts out and revises or develops in a back-and-forth process, are simply one's credences. Credences are a paradigmatically degreed type of propositional attitude, in that they are associated with real numbers or intervals between 0 and 1 (inclusive). I have also suggested that the outputs of reflective equilibrium, i.e. the 'theories' with which one ends up, are acceptances in Cohen's sense, i.e. a matter of having a policy of including the relevant proposition among one's premises in some context (e.g. for the purposes of understanding something). And I've suggested that this fits well with the common practice of assuming that the outputs of reflective equilibrium are binary rather than a matter of degree. So, on the view I've sketched thus far, the propositional attitudes involved in having 'judgments' and 'theories' are of quite different sorts, viz. credences and acceptances. The question remains, of course, how to move from one to the other. By what process can we get from (degreed) credences to (binary) acceptances? As noted in the introduction, however, I focus in this paper on the more modest issue of how to conceive of the *state* of reflective equilibrium. My central concern, therefore, will be how to account for the relationship between credences and acceptances at that final stage in the process, i.e. when (the state of) reflective equilibrium has been reached.

4 A probabilistic model of reflective equilibrium

The suggestion I wish to explore in the rest of this paper gives a central role to probability. In particular, probability enters the picture through the assumption that in order for an agent to come to be in a state of reflective equilibrium, the agent's credences—which may initially have taken any values whatsoever—must at that point have become *probabilistically coherent*, i.e. so as to satisfy the axioms of the probability calculus. To motivate this assumption, note that the agent whose credences fail to be probabilistically coherent in this sense would fail to be epistemically rational according to several powerful arguments. For instance, arguments that appeal to epistemic utility

[11] As Roorda (1997, pp. 148–149) elegantly puts it, "we do not require the gambler to make sure that all of the propositions he bets on be logically consistent; but we do require of the storyteller that the logical consequences of what she has already said will not be contradicted as the story unfolds." Thus, assuming that we require as much of the philosopher as we do of the storyteller, reflective equilibrium evidently involves binary belief-like attitudes, e.g. of the sort Cohen calls acceptance.

theory purport to show that the credences of an agent who fails to be probabilistically coherent would, by the agent's own lights, be more likely be accurate if their credences were to be modified so as to become probabilistically coherent (Joyce, 1998; Pettigrew, 2016).[12] Whatever else being in reflective equilibrium involves, it seems that it at least requires that one's credences aren't transparently irrational in this way.

So an agent's credences, in reflective equilibrium, must evidently be probabilistically coherent. However, this is at most half of the story about what it is for someone's attitudes to be in reflective equilibrium at a given time. For as I have argued above, the output of the method of reflective equilibrium, i.e. 'theories', are claims that the agent *accepts*, where acceptance is a binary attitude that is quite distinct from credence. Hence the crucial question, for our purposes, is what propositions may or should be accepted by an agent in reflective equilibrium—that is, by an agent that has various credences, which can at this point be assumed to be probabilistically coherent. In brief, the question is: Given some (probabilistically coherent) credences in propositions $P_1, ..., P_n$, which (if any) of the P_is would an agent in reflective equilibrium also accept (or be prepared to accept)?

4.1 The threshold model

A natural view is that an agent in reflective equilibrium accepts (or is prepared to accept)[13] a proposition P_i just in case her credence—or, equivalently at this point, her subjective probability[14]—is equal to or greater than some threshold value $t \in (0, 1)$.[15] Unfortunately, however, this view quickly leads to paradoxical results.[16] For suppose an agent assigns a probability at or above t to two probabilistically independent claims P_i and P_j. Since the probability of their conjunction $P_i \wedge P_j$ is necessarily lower than the probability of each conjunct, the former probability will be below t in many cases of this sort. To illustrate, suppose we set our threshold at $t = 0.75$. Then take two probabilistically independent propositions P_i and P_i such that $Pr(P_i) = Pr(P_j) = 0.8$, for instance. The probability axioms imply that, in such a case, $Pr(P_i \wedge P_j) = 0.8^2 = 0.64$. So the view we are currently considering would imply that an agent in reflective equilibrium accepts each of two propositions separately, but simultaneously does not accept their conjunction—even though the latter obviously follows deductively from the former. This is a problematic consequence of the view for at least two separate reasons:

First, it is not clear that any sense can be made of the idea of accepting P_i and P_j separately but not their conjunction $P_i \wedge P_j$. Recall that, by our Cohen-inspired

[12] Other influential classes of arguments for this conclusion include synchronic Dutch Book arguments (Vineberg, 2022, §2) and arguments that appeal to representation theorems (Titelbaum, 2022, Chap. 8).

[13] In what follows, I will sometimes omit this parenthetical qualification and just formulate the view as claiming something about what an agent accepts in reflective equilibrium. This should be taken as shorthand for the more precise formulation here, however.

[14] An agent's subjective probabilities are simply the agent's probabilistically coherent credences.

[15] Analogous views have been proposed and defended concerning outright belief under the label 'the Lockean thesis' (see, e.g., Foley, 1992; Sturgeon, 2008).

[16] The paradoxical results are analogous in structure to the well-known lottery and preface paradoxes (Kyburg, 1961; Makinson, 1965).

definition of 'acceptance', to accept a proposition is to have a policy of including it among one's premises for deciding what to do or think in a particular context (e.g., in the context of understanding something). But what would it be to include both P_i and P_j, but not $P_i \wedge P_j$, among one's premises for deciding what to do or think in a particular context? By including P_i and P_j, one would seem to be effectively including $P_i \wedge P_j$ as well (and *vice versa*), since any decision regarding what to do or think made on the basis of the former premises can be made on the basis of the latter (and *vice versa*). So it seems that there can be no daylight between accepting P_i and P_j, on the one hand, and accepting $P_i \wedge P_j$, on the other, contrary to what the view we are now considering would have to suppose.

Second, even if we could somehow make sense of the idea of accepting each of two claims but not their conjunction, there remains the issue of whether agents in reflective equilibrium would be required to instantiate such a paradoxical combination of attitudes. At least when it comes to philosophical theorizing, it is generally considered better to have a single, comprehensive theory than having several distinct theories—and this seems no less true of the sort of theorizing that occurs under the banner of reflective equilibrium. So it is hard to see why reflective equilibrium would require, on purely formal grounds, accepting each of P_i and P_j but not their more comprehensive combination $P_i \wedge P_j$. Furthermore, note that a given proposition P that isn't sufficiently probable to exceed the threshold t (but whose probability isn't zero) could generally be 'broken up' into a conjunction of several distinct propositions $P_1, ..., P_n$ (where $P \equiv P_i, ..., P_n$) such that the probability of each proposition P_k in this series exceeds t simply in virtue of being logically weak. On the view we are now considering, each such P_k would be accepted in reflective equilibrium, while P itself could not be accepted. So, for example, on this view an agent in reflective equilibrium might have to accept every single implication of utilitarianism at a certain level of granularity (e.g. regarding what to do in various concrete situations) while also being required not to accept utilitarianism as whole. This is absurd.

I conclude, then, that there is no hope for the view that an agent in reflective equilibrium accepts a proposition just in case its probability exceeds some threshold t. Since I have not specified any value for the threshold t in my arguments against this view, it should be clear that this rules out *any* view which imposes such a threshold for acceptance in reflective equilibrium—even one that varies with the context of utterance or what's at stake for the agent, for example.[17] Put differently, the entire approach of accounting for acceptance in reflective equilibrium by appealing to a probability threshold is on the wrong track. Fortunately, there is another way in which probabilities may come into play in an account of acceptance in reflective equilibrium.

4.2 The optimality model

To motivate this alternative view in general terms, consider the distinction between satisficing and optimizing introduced in a different context by the economist Herbert

[17] Views in which what an agent knows, believes, or is justified in believing depends on such factors have been defended under the general label of 'contextualism' in recent years (Clarke, 2011; DeRose, 1992; Rysiew, 2021).

Simon (1956). In *satisficing*, one chooses an option that is sufficiently good, i.e. satisfactory, according to some metric. In *optimizing*, one chooses an option that is the best of all alternatives, i.e. optimal, according to some metric. The metric can be the same in both cases. In particular, if the metric is probability, and the question is whether or not to accept some theory, then satisficing amounts to accepting all and only theories that are sufficiently probable, i.e. that exceed some threshold probability; whereas optimizing amounts to accepting any theory that is more probable than its alternatives. It should be clear that the satisficing approach to reflective equilibrium is a non-starter, for it leads directly to the threshold view that we have just found wanting. So let us instead consider whether an optimizing approach might fare better.

It is far from obvious, however, how to develop an optimizing approach to reflective equilibrium without running into problems that are as serious as those that face the threshold view. In particular, what are the 'alternatives' that are being compared probabilistically? And how can we ensure that the 'best' of these alternatives is not simply the least ambitious theory, which will generally be more probable than its more ambitious counterparts? Finally, how can we ensure that the set of theories that an agent accepts or is prepared to accept in reflective equilibrium will be deductively consistent and closed under deductive consequence? Happily, I have in previous work on acceptance and understanding already developed a detailed model of rational acceptance which is designed to deal with issues of this kind, viz. what I call *the Optimality Model* (Dellsén, 2021; see also Dellsén, 2018). A crucial feature of this model is that the rational acceptability of different theories is determined *holistically*, i.e. roughly in terms of how they fit into a much larger theoretical corpus. In this model, a given proposition is rationally acceptable in the context of understanding something just in case it follows from a maximally informative theory of the understood phenomenon that is more probable, or much more probable, than any alternative such account.

A key element of the Optimality Model is the notion of a 'comprehensive theory'. A theory T^+ is defined to be *comprehensive*, relative to a set of an agent S's questions Q, if and only if, for any question $q_k \in Q$, T^+ provides a complete answer to q_k. Here, the notion of a 'complete answer' to a question q is defined, relative to a set of possible answers $\{A_1, \ldots, A_n\}$ to q, as a conjunction of n propositions in which the i-th member is either A_i or $\neg A_i$. So to say that T^+ is comprehensive, relative to a set of questions Q, is to say that it implies such a conjunction of possible answers, or their negations, to any question in Q. For example, as discussed further below, a maximally specific version of utilitarianism would be a comprehensive theory relative to a set of questions regarding morally correct behavior if it implies that any given answer to those questions is either true or false. For instance, such a version of utilitarianism would presumably imply various answers to the question 'On what does the moral correctness of an action depend?', such as 'Its consequences' and 'Which other actions were available to the agent', as well as the negations of various other answers, such as 'The agent's intentions' and 'Which rule the agent was following'.

With this characterization of 'comprehensive theory' in place, an extension of the Optimality Model to reflective equilibrium, slightly streamlined with some minor modifications to the original model, can be succinctly stated as follows:

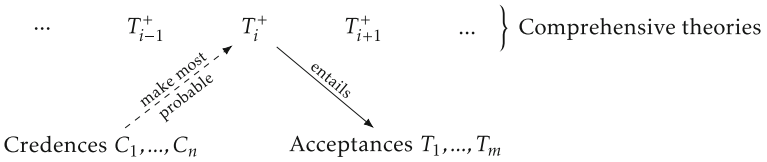

Fig. 1 A schematic diagram of the relationship between credences ('judgments') and acceptances ('theories') in reflective equilibrium, according to OMRE

> **The Optimality Model of Reflective Equilibrium (OMRE):** An agent S is in reflective equilibrium just in case (i) S's credences are probabilistically coherent, and (ii) any theory T that S accepts (or is prepared to accept) is entailed by a 'comprehensive' theory T^+ that is [much] more probable, by the lights of S's credences, than any alternative such theory [that entails $\neg T$].

I will not here go into the various technical details of this model, such as how the model can be tweaked with the bracketed option of requiring T^+ to be *much* more probable than its alternatives (which requires a comparison between T^+ and the most probable comprehensive theory that entails $\neg T$). (For such details, the interested reader may consult Dellsén 2021, pp. 2485–2489.) Rather, I will focus on conveying the general idea behind the model and what it means for how the notion of reflective equilibrium is used in philosophical argumentation through informal explanations, examples, and a diagram (see Fig. 1).

The idea behind OMRE is that for a given topic or phenomenon there will be a set of *comprehensive* alternative theories in logical space that would answer any question one has regarding that topic or phenomenon. To be clear, these comprehensive theories need not have actually been formulated or even considered by the agent in question; but there will nevertheless *exist* (in logical space) a set of such comprehensive theories. For example, I take it that fully fledged out versions of utilitarianism, deontology, and virtue ethics, are comprehensive theories relative to a set of questions regarding morally correct behavior, in so far as each such fully fledged out theory purports to answer any question that one has regarding morally correct behavior. (Note that I am referring here to *fully fledged out versions* of utilitarianism, deontology and virtue ethics; these will presumably be a great deal more informative than what usually gets referred to as 'utilitarianism', 'deontology' and 'virtue ethics' in so far as those theories do not literally answer *all* of one's questions concerning morally correct behavior. Indeed, it may well be that no ethicist has ever formulated a fully comprehensive theory of normative ethics, because each actually formulated ethical theory addresses only some limited range of the questions regarding morally correct behavior.)

In any case, there will normally be one and only one of these comprehensive theories that is most probable by the lights of the credences that constitute an agent's 'judgments' in reflective equilibrium.[18] This particular comprehensive theory will have numerous implications, both particular and general. For example, a particular

[18] What if there are multiple comprehensive theories that are exactly equally probable, but each more probable than the remaining comprehensive theories? OMRE, as formulated above, implies that in such a case, no theory may be accepted by S, since there is then no comprehensive theory that is more probable than *any* alternative comprehensive theory. It is unclear to me what to say about such cases, but if one finds

fully fledged out version of utilitarianism—some version of hedonistic, actualist, act-utilitarianism, let's say—might imply that one ought to pull the lever in the original trolley case so as to cause the death of one person but thereby save five others (a particular claim), and also that only the consequences of one's actions are morally relevant (a general claim). If that comprehensive theory is, by the lights of one's subjective probability assignments, more probable than any other comprehensive theory (e.g., another fully fledged out version of utilitarianism, or a fully fledged out version of some non-utilitarian theory) then an agent in reflective equilibrium would accept (or be prepared to accept) any of these implicated claims—including, of course, the comprehensive theory itself—according to OMRE.

So, in sum, OMRE holds that in an agent in reflective equilibrium accepts a given theory T just in case it is part and parcel of what is by one's own lights the most probable comprehensive theory, i.e. a theory that is sufficiently strong so as to answer any question one has regarding the phenomenon or topic in question. If this seems painfully obvious or even trivial, it is perhaps worth highlighting what this picture does *not* require of an accepted theory T. In particular, this does not require that the probability of T exceeds any threshold greater than zero. Indeed, there are certain cases in which such a T would be very improbable by the agent's own lights, because although T is not probable in and of itself, it is an indispensable part of a comprehensive theory that is more probable than any other.[19] Such is the holistic nature of OMRE. In the next section, we will see why this holism—which may initially seem like a liability—is in fact one of the model's several virtues.

5 Getting what we want from reflective equilibrium

In Sect. 2, I suggested that an account of reflective equilibrium should ideally satisfy three desiderata: (i) it should make sense of the ways in which philosophers who appeal to reflective equilibrium argue for their theories; (ii) it should fit or cohere with a general epistemological framework that is independently motivated; and (iii) it should imply that accepted theories are consistent with one another and that acceptance can be extended to any deductive implication of what one already accepts. Let us now go through these desiderata in reverse order and discuss how the Optimality Model of Reflective Equilibrium (OMRE) handles each desideratum.

5.1 Revisiting deductive cogency of theories

Consider first the desideratum that the set of theories one accepts in reflective equilibrium should be deductively cogent, i.e. consistent and closable (see Sect. 2.3). This is very much a non-trivial desideratum within a probabilistic framework, because as we have effectively seen already in our discussion of the probability threshold model (see

OMRE's verdict unsatisfactory here one may easily modify OMRE so as to allow a theory T be accepted in such cases if T is entailed by *each* of the multiple equally probable comprehensive theories.

[19] As Folke Tersman has pointed out to me in private communication, this feature of OMRE closely resembles some of Scheffler's (1954, pp. 181–183) remarks about justification in ethics, which were in turn inspired by Goodman (1952, p. 163).

Sect. 4.1), there are well known obstacles to combining probabilistic requirements on acceptance with the requirement that accepted theories be deductively cogent. In particular, as long as one takes a satisficing approach, thus effectively imposing some sort of probability threshold on acceptance, there is no way to satisfy the requirement that accepted theories be deductive cogent, because there will always be propositions that are individually more probable than the threshold and yet are either inconsistent with one another or so as to imply another proposition that falls below the threshold. Either way, deductive cogency is violated.

Happily, such violations of deductive cogency are avoided if one takes the optimizing approach suggested by OMRE, for one can show (see Dellsén, 2021, pp. 2489–2491) that a probabilistic model of acceptability that shares all the formal properties of OMRE validates deductive cogency. So, in effect, OMRE provably implies that the set of theories that one accepts in reflective equilibrium will be deductively consistent and closable under deductive consequence. For example, if OMRE implies that an agent S in reflective equilibrium accepts (or is prepared to accept) a given set of theories $\{T_1, ..., T_m\}$, then (a) $\{T_1, ..., T_m\}$ will be consistent, and (b) S also accepts (or is prepared to accept) any T_k that is entailed by some or all of $\{T_1, ..., T_m\}$, including the conjunction $T_1 \wedge \cdots \wedge T_m$.[20] This is a significant result, not least because these implications of the OMRE are not baked into the model by ad hoc stipulation, e.g. by explicitly imposing a special deductive cogency constraint on reflective equilibrium; rather, they fall naturally out of the core idea that the claims accepted in reflective equilibrium are those implied by the optimally probable comprehensive theory.

5.2 Revisiting fit with general epistemology

Consider next the desideratum that our account of reflective equilibrium should cohere or fit with an independently-motivated general epistemological framework, as opposed to standing out as an ad hoc exception to such a general framework (see Sect. 2.2). It should be clear that OMRE measures up to this desideratum. OMRE is simply an *extension* of the approach to epistemology known variously as Bayesianism or Probabilism, albeit without committing to some of the more controversial aspects that are sometimes associated with these labels (such as particular updating rules, e.g. Bayesian Conditionalization). Although the formal methods typically employed by those who work within this framework may not be to everyone's taste, it's hard to deny that the Bayesian framework has proved to be remarkably successful in illuminating a huge variety of epistemic phenomena (see, e.g., Lin, 2022, §§1.3 & 7).

Indeed, it is worth adding that the Bayesian framework also coheres especially nicely with the dominant normative account of rational choice, viz. *expected utility*

[20] To get an intuitive sense of why (b) holds, for example, consider that OMRE dictates that for each $T_1, ..., T_m$ to be acceptable, there must be an optimally probable comprehensive theory T^+ which entails them all. Since entailment is transitive, T^+ would then also entail anything that is entailed by $T_1, ..., T_m$ (jointly or separately), including the conjunction $T_1 \wedge \cdots \wedge T_m$. Hence any such entailment from $T_1, ..., T_m$ would be included amongst the theories that S accepts, or is prepared to accept, in reflective equilibrium according to OMRE.

theory, which appeals centrally to subjective probabilities (Briggs, 2019).[21] In fact, it also coheres very nicely with most of the most prominent rivals to that account, which also appeal to subjective probabilities (e.g., Buchak, 2022; Steele & Stefánsson, 2020)! Relatedly, it also coheres nicely with widely accepted descriptive theories of decision making, reasoning, and the mind more generally, that have been developed within the fields of economics, psychology, and cognitive science (see, e.g., Greenberg, 2013; Knill & Pouget, 2004). So OMRE actually goes further than the original desideratum demanded, in not only cohering with an independently-plausible general epistemological framework, but also with various widely endorsed theories in other disciplines and subdisciplines that are entangled in various ways with epistemology and philosophy in general.

5.3 Revisiting philosophical argumentation

Finally, consider the desideratum that our account of reflective equilibrium should make sense of the ways in which philosophers who appeal to reflective equilibrium argue for their theories (see Sect. 2.1). Recall, though, that we immediately qualified this by noting that our model need not imply that every instance in which philosophers appeal to reflective equilibrium is a legitimate instance thereof, since individual philosophers may be confused about what reflective equilibrium does or should be taken to require. However, certain core features of philosophical argumentation which explicitly appeal to reflective equilibrium should arguably be explained by any adequate account thereof. In a way, I have already discussed a few such features above, such as the requirement that accepted theories be deductively consistent and closable. In what follows, I will thus only comment on two additional features of this sort that I take to be highly significant.

The first such feature is that reflective equilibrium should involve a *holistic* rather than piecemeal evaluation of the theories we end up accepting.[22] I take this to mean that a particular theory might be accepted in virtue of being an indispensable part of a more general theory (or set of theories) that is (are) plausible or well-supported when considered as a whole, even if that particular claim would seem implausible or poorly supported when viewed in isolation. For example, most theories in normative ethics have implications for particular cases that are viewed as implausible in isolation, even by their own proponents. However, these proponents are prepared to 'bite the bullet' on those particular claims because the general theory is deemed to be sufficiently plausible or well-supported when viewed from a holistic perspective—which includes an evaluation not only of the theory as a whole, but also of how that entire theory fits with various background theories, e.g. from other domains of inquiry.[23]

[21] As I explain in detail below (see Sect. 5.3), OMRE does not in any way require agents to revise their credences in light of what they accept, e.g. by setting the credences in accepted propositions to 1.

[22] That reflective equilibrium involves an holistic evaluation of theories is a point made explicitly by, among others, Nielsen (1994), Føllesdal (2005), Elgin (2017), and Baumberger and Brun (2021). It is also implicit in the writings of, e.g., Goodman (1952), Rawls (1999), and Daniels (1979).

[23] That reflective equilibrium should be taken to involve evaluating theories holistically also with respect to background theories from other domains is a point influentially made by Daniels (1979), who labels the resulting idea 'wide reflective equilibrium'.

OMRE provides a particularly satisfying explanation of this holistic nature of reflective equilibrium. As mentioned at the end of the previous section (Sect. 4.2), a particular bit of theory T may be accepted by an agent S in reflective equilibrium on this model even if S has a very low rational credence in T—provided that T is implied by a comprehensive theory T^+ that is more probable by S's lights than any other. In this way, T inherits its acceptability from the acceptability of the more general, comprehensive theory T^+, as opposed to being deemed acceptable in isolation from other related claims. Thus, from the point of view of OMRE, it is entirely appropriate to 'bite the bullet' on pre-theoretically implausible implications of one's theories by accepting them in reflective equilibrium while nevertheless having a low credence in such claims. In this sense, our 'intuitions' can be overridden by the theories we end up accepting in reflective equilibrium.

However, it is worth noting that OMRE does not require agents to *revise* their credences in a theory T once they have accepted T; rather, as far as OMRE is concerned, an agent's credences remain unaffected by what the agent accepts.[24] To be sure, one *could* add such a requirement to OMRE, e.g. by postulating that once our agent S has accepted T, she then must raise her credence in T to 1, or perhaps some other threshold level t. However, adding such a requirement would arguably be both unmotivated and problematic. It would be unmotivated because there is no obvious reason why an agent's credences ought to be change in light of what she accepts in this way once we have given up on the idea of linking acceptance to subjective probability above some threshold (see Sect. 4.1). An agent may accept a theory T, i.e. include it among the premises for deciding what to do or think in some context, and yet at the same time find T implausible or unlikely to be true.[25] Furthermore, requiring that agents in reflective equilibrium revise their credences in light of what they accept is problematic because it would imply that agents should systematically update their credences in ways that conflict with Bayesian Conditionalization (and generalizations thereof, such as Jeffrey Conditionalization). This would in turn make agents in reflective equilibrium immediately susceptible to dynamic Dutch Books (see Vineberg, 2022, §4) and systematically irrational according to related arguments based on epistemic utility theory (e.g., Greaves & Wallace, 2006).

In sum, then, I see no real benefits, but some serious problems, with adding a requirement to the effect that an agent who accepts T must then revise or raise her credence in T. Rather, credence and acceptance should simply be seen as different propositional attitudes that are normatively related only in that T is accepted in reflective equilibrium just in case it follows from the comprehensive theory that is most probable in light of those credences.

Another feature of philosophical argumentation that appeals to reflective equilibrium concerns philosophers' frequent appeals to explanatory virtues. As noted above

[24] This is not to say that the *method* of reflective equilibrium leaves an agent's credences unchanged. As noted above (see Sect. 4; see also Sect. 3.1), an agent's credences will typically have to become probabilistically coherent during the process described by the method, which corresponds to the common idea that reflective equilibrium involves a back-and-forth process of revising judgments in light of each other.

[25] For direct arguments to this effect, see Wilkenfeld (2016), Baumberger (2018), and my own previous work (Dellsén, 2017, 2018). More generally, this also follows from Cohen's definition of acceptance, which is explicitly contrasted with belief and degrees thereof (Cohen, 1992, pp. 1–27).

(Sect. 2.1), philosophers often assume that, all other things being equal, a simpler theory should be preferred to a more complex one, a unifying theory should be preferred to a collection of separate theories, and so on for various other explanatory virtues. Happily, there is a whole literature in philosophy of science that is specifically concerned with explaining how preferences for explanatorily virtuous theories arise naturally within the Bayesian framework (e.g., Henderson, 2014; Huemer, 2009; McGrew, 2003). In short, these authors show that, given plausible assumptions about common credence assignments, Bayesian agents will assign probabilities to theories that exhibit the various explanatory virtues, e.g. simplicity and unification. Given this, OMRE not only allows for explanatory virtues to play an important role in reflective equilibrium, but also *explains* why and how they play this role.

For example, there are several reasons within the Bayesian framework for thinking that, other things being equal, more unified theories are more likely to be true (Blanchard, 2018; Lange, 2004; McGrew, 2003; Myrvold, 2003; Schupbach, 2005). In light of these considerations, OMRE explains why the theories accepted in reflective equilibrium should generally be unified. After all, the comprehensive theories that entail any accepted theory would generally have to be at least somewhat unified in order to be more probable than all alternative such theories; otherwise, their probability would suffer as compared to their more unified alternatives (all other things being equal). Put differently, a high degree of unification would be a common by-product of optimal probability—and thus function as reliable indicator of optimal probability. In this way, OMRE would not need to impose separate ad hoc requirements in order to account for the preference for explanatorily virtuous theories; rather, the preference for explanatorily virtuous theories falls out of the core elements of the model without postulating any additional requirements on reflective equilibrium.

6 Conclusion

This paper has aimed to provide a model of the state of reflective equilibrium within a probabilistic framework for epistemic rationality. The key idea behind the model is that an agent in reflective equilibrium accepts, or is prepared to accept, all and only claims that follow from a maximally informative theory that is more probable by the agent's lights than any other such theory. Drawing on previous work, we have seen that when an agent is in reflective equilibrium thus understood, the set of claims they accept or are prepared to accept is bound to be deductively consistent and closed under deductive implication. Furthermore, we have seen that this model can explain various features of philosophical argumentation in which the notion of reflective equilibrium features centrally, such as why philosophical theories are evaluated holistically in reflective equilibrium, and why some 'intuitive' claims may be overridden by the theories we end up accepting in reflective equilibrium.

As I have emphasized throughout, this model of reflective equilibrium does not purport to capture all aspects of what has become known as 'reflective equilibrium'. In particular, I have had little to say about the step-by-step process of bringing one's attitudes into the *state* of reflective equilibrium that is often taken to be an important part of the *method* of reflective equilibrium—although I am hopeful that the current

model can be extended to cover this process as well (perhaps building on work by, e.g., Staffel, 2019). I have also mostly kept the discussion at a high level of generality and abstraction, focusing on big picture issues rather than concrete applications of reflective equilibrium (in contrast to, e.g., Rechnitzer, 2022)—but, again, I am optimistic that the model presented here could be fruitfully applied to concrete situations in future work.

Acknowledgements This paper was first presented at the conference *Reflective Equilibrium 51 Years after A Theory of Justice* at the University of Bern in the spring of 2022. A am grateful to the audience there, especially Georg Brun, Catherine Elgin, and Folke Tersman, for very helpful discussions that helped improve the paper a great deal. I am also very grateful to the editors of this special issue, three anonymous reviewers, and to Folke Tersman, for written feedback that led to various significant improvements. My research for paper was supported by Icelandic Research Fund Grant 228526-051, *Understanding Philosophical Progress*.

Funding Open access funding provided by Inland Norway University Of Applied Sciences.

Declarations

Conflict of interest I confirm that I have no conflicts of interest to report.

References

Baumberger, C. (2018). Explicating objectual understanding: Taking degrees seriously. *Journal for General Philosophy of Science, 50*(3), 367–388.

Baumberger, C., & Brun, G. (2021). Reflective equilibrium and understanding. *Synthese, 198*, 7923–7947.

Blanchard, T. (2018). Bayesianism and explanatory unification: A compatibilist account. *Philosophy of Science, 85*(4), 682–703.

Briggs, R. A. (2019). Normative theories of rational choice: Expected utility. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2019 edition). Metaphysics Research Lab, Stanford University.

Brun, G. (2014). Reflective equilibrium without intuitions? *Ethical Theory and Moral Practice, 17*, 237–252.

Buchak, L. (2022). Normative theories of rational choice: Rivals to expected utility. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2022 edition). Metaphysics Research Lab, Stanford University.

Christensen, D. (2004). *Putting logic in its place: Formal constraints on rational belief*. Oxford University Press.

Clarke, R. (2011). *Belief in context*. PhD thesis, The University of British Columbia, Vancouver.

Cohen, L. J. (1992). *An essay on belief and acceptance*. Clarendon Press.

Daniels, N. (1979). Wide reflective equilibrium and theory acceptance in ethics. *Journal of Philosophy, 76*, 256–282.

De Bona, G., & Staffel, J. (2018). Why be coherent? *Analysis, 78*, 405–415.

Dellsén, F. (2017). Understanding without justification or belief. *Ratio, 30*, 239–254.

Dellsén, F. (2018). Deductive cogency, understanding, and acceptance. *Synthese, 195*, 3121–3141.
Dellsén, F. (2020). Beyond explanation: Understanding as dependency modeling. *The British Journal for the Philosophy of Science, 71*, 1261–1286.
Dellsén, F. (2021). Rational understanding: Toward a probabilistic epistemology of acceptability. *Synthese, 198*, 2475–2494.
DeRose, K. (1992). Contextualism and knowledge attributions. *Philosophy and Phenomenological Research, 52*, 913–929.
Elgin, C. Z. (1996). *Considered judgment*. Princeton University Press.
Elgin, C. Z. (2004). True enough. *Philosophical Issues, 14*, 113–131.
Elgin, C. Z. (2006). From knowledge to understanding. In S. Hetherington (Ed.), *Epistemology futures* (pp. 199–215). Oxford University Press.
Elgin, C. Z. (2007). Understanding and the facts. *Philosophical Studies, 132*, 33–42.
Elgin, C. Z. (2009). Is understanding factive? In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Epistemic value* (pp. 322–330). Oxford University Press.
Elgin, C. Z. (2017). *True enough*. MIT Press.
Foley, R. (1992). The epistemology of belief and the epistemology of degrees of belief. *American Philosophical Quarterly, 29*, 111–124.
Føllesdal, D. (2005). The emergence of justification in ethics. *European Review, 13*, 169–182.
Goodman, N. (1952). Sense and certainty. *The Philosophical Review, 61*, 160–167.
Goodman, N. (1955). *Fact, fiction, and forecast*. Harvard University Press.
Greaves, H., & Wallace, D. (2006). Justifying conditionalization: Conditionalization maximizes expected epistemic utility. *Mind, 115*, 607–632.
Greenberg, E. (2013). *Introduction to Bayesian econometrics*. Cambridge University Press.
Grimm, S. (2006). Is understanding a species of knowledge? *British Journal for the Philosophy of Science, 57*, 515–535.
Grimm, S. (2014). Understanding as knowledge of causes. In A. Fairweather (Ed.), *Virtue epistemology naturalized: Bridges between virtue epistemology and philosophy of science* (pp. 347–360). Springer.
Henderson, L. (2014). Bayesianism and Inference to the best explanation. *British Journal for the Philosophy of Science, 65*, 687–715.
Hills, A. (2016). Understanding why. *Nous, 50*, 661–688.
Huemer, M. (2009). When is parsimony a virtue. *Philosophical Quarterly, 59*, 216–236.
Jäger, C., & Malfatti, F. I. (2021). The social fabric of understanding: Equilibrium, authority, and epistemic empathy. *Synthese, 199*, 1185–1205.
Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science, 65*, 575–603.
Kaplan, M. (1996). *Decision theory as philosophy*. Cambridge University Press.
Kaplan, M. (2013). Coming to terms with our human fallibility: Christensen on the preface. *Philosophy and Phenomenological Research, 87*, 1–35.
Kauppinen, A., & Hirvelä, J. (2023). Reflective equilibrium. In D. Copp, T. Rulli, & C. Rosati (Eds.), *The Oxford handbook of normative ethics.* Oxford University Press.
Kelly, T., & McGrath, S. (2010). Is reflective equilibrium enough? *Philosophical Perspectives, 24*, 325–359.
Kelp, C. (2017). Towards a knowledge-based account of understanding. In S. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding: Perspectives from epistemology and philosophy of science* (pp. 251–271). Routledge.
Khalifa, K. (2017). *Understanding, explanation, and scientific knowledge*. Cambridge University Press.
Kim, J. (1974). Noncausal connections. *Nous, 8*, 41–52.
Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27*(12), 712–719.
Kvanvig, J. (2003). *The value of knowledge and the pursuit of understanding*. Cambridge University Press.
Kyburg, H. E. (1961). *Probability and the logic of rational belief*. Wesleyan University Press.
Lange, M. (2004). Bayesianism and unification: A reply to Wayne Myrvold. *Philosophy of Science, 71*, 205–215.
Lin, H. (2022). Bayesian epistemology. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2022 edition). Metaphysics Research Lab, Stanford University.
Lycan, W. G. (1985). Epistemic value. *Synthese, 64*, 137–164.
Makinson, D. C. (1965). The paradox of the preface. *Analysis, 25*, 205–207.
McGrew, T. (2003). Confirmation, heuristics, and explanatory reasoning. *British Journal for the Philosophy of Science, 54*, 553–567.

McPherson, T. (2015). The methodological irrelevance of reflective equilibrium. In C. Daley (Ed.), *Palgrave handbook of philosophical methodology* (pp. 652–674). Palgrave Macmillan.
Mizrahi, M. (2012). Idealizations and scientific understanding. *Philosophical Studies, 160*, 237–252.
Myrvold, W. C. (2003). A Bayesian account of the virtue of unification. *Philosophy of Science, 70*(2), 399–423.
Nielsen, K. (1994). Philosophy within the limits of wide reflective equilibrium alone. *The Jerusalem Philosophical Quarterly, 43*, 3–41.
Pettigrew, R. (2016). *Accuracy and the laws of credence*. Oxford University Press.
Pritchard, D. (2009). Knowledge, understanding, and epistemic value. In A. O'Hear (Ed.), *Epistemology (Royal Institute of Philosophy Lectures)* (pp. 19–43). Cambridge University Press.
Rawls, J. (1999). *A theory of justice* (Revised). Harvard University Press.
Rechnitzer, T. (2022a). *Applying reflective equilibrium: Towards the justification of a precautionary principle*. Springer.
Rechnitzer, T. (2022b). Turning the trolley with reflective equilibrium. *Synthese, 200*, 272.
Roorda, J. (1997). Fallibilism, ambivalence, and belief. *Journal of Philosophy, 94*, 126–155.
Rysiew, P. (2021). Epistemic contextualism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2021 edition). Metaphysics Research Lab, Stanford University.
Scheffler, I. (1954). On justification and commitment. *The Journal of Philosophy, 51*, 180–190.
Schupbach, J. N. (2005). On a Bayesian analysis of the virtue of unification. *Philosophy of Science, 72*, 594–607.
Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*, 129–138.
Singer, P. (1974). Sidgwick and reflective equilibrium. *The Monist, 58*, 490–517.
Sliwa, P. (2015). Understanding and knowing. *Proceedings of the Aristotelian Society, 115*, 57–74.
Staffel, J. (2019). *Unsettled thoughts: A theory of degrees of rationality*. Oxford University Press.
Steele, K., & Stefánsson, H. O. (2020). Decision theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2020 edition). Metaphysics Research Lab, Stanford University.
Sturgeon, S. (2008). Reason and the grain of belief. *Noûs, 42*, 139–165.
Tersman, F. (1993). *Reflective equilibrium: A essay in moral epistemology*. Almqvist & Wiksell.
Titelbaum, M. (2022). *Fundamentals of Bayesian epistemology 2: Arguments, challenges, alternatives*. Oxford University Press.
Vallinder, A. (2018). *Bayesian variations: Essays on the structure, object, and dynamics of credence*. PhD thesis, London School of Economics and Political Science.
van Fraassen, B. C. (1995). Fine-grained opinion, probability, and the logic of full belief. *Journal of Philosophical Logic, 24*, 349–377.
Vineberg, S. (2022). Dutch book arguments. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2022 edition). Metaphysics Research Lab, Stanford University.
Weisberg, J. (2020). Belief in psyontology. *Philosopher's Imprint, 20*(11), 1–27.
Wilkenfeld, D. A. (2016). Understanding without believing. In S. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding: Perspectives from epistemology and philosophy of science* (pp. 318–334). Routledge.