Synthese (2020) 197:925–954
https://doi.org/10.1007/s11229-017-1596-4

S.I. : PHILMETHODS

Conceptual re-engineering: from explication to reflective equilibrium

Georg Brun[1]

Received: 9 February 2017 / Accepted: 15 October 2017 / Published online: 27 October 2017

Abstract Carnap and Goodman developed methods of conceptual re-engineering known respectively as explication and reflective equilibrium. These methods aim at advancing theories by developing concepts that are simultaneously guided by pre-existing concepts and intended to replace these concepts. This paper shows that Carnap's and Goodman's methods are historically closely related, analyses their structural interconnections, and argues that there is great systematic potential in interpreting them as aspects of one method, which ultimately must be conceived as a component of theory development. The main results are: an adequate method of conceptual re-engineering must focus not on individual concepts but on systems of concepts and theories; the linear structure of Carnapian explication must be replaced by a process of mutual adjustments as described by Goodman; Carnap's condition of similarity can be analysed into two components, one securing a relation to the specific extensions of the pre-existing concepts, one regulating the transition to the new system of concepts; these two criteria of adequacy can be built into Goodman's account of reflective equilibrium to ensure that the resulting concepts promote theoretical virtues while being sufficiently similar to the concepts we started out with.

Keywords Explication · Reflective equilibrium · Definition · Reconstruction · Conceptual engineering · Carnap · Goodman

✉ Georg Brun
Georg.Brun@philo.unibe.ch

[1] Institute for Philosophy, University of Bern, Länggassstrasse 49a, 3000 Bern 9, Switzerland

1 Introduction

Philosophers and scientists develop concepts to solve theoretical and practical problems. Examples range from newly coined terms such as “speciesism” in ethics and “quark” in physics, the introduction of new uses for old words such as “ring” in mathematics and “frame” in modal logic, to re-definitions of concepts in use, for example, *terrorist*[1] in political philosophy, *work* in economics and *planet* in astronomy. If done explicitly and intentionally, such development of concepts can be called “conceptual engineering”; if it is guided by a concept in use and simultaneously aims at replacing this concept, we may speak of “conceptual re-engineering”. In philosophy, as in the sciences, we are especially interested in conceptual re-engineering that serves the purpose of advancing theory. Examples can be found in all philosophical disciplines and epochs. They include Tarski’s ‘definitions’ of *truth* (1983) and *logical consequence* (2002), Hempel’s explication of *explanation* (e.g. 1970), Kant on *opinion*, *belief* and *certainty* (1998: p. A 822/B 850), Rawls’s explication of *justice* and *rightness* (1999c:esp. p. 96), Griffin’s notion of *human right* (2008) and Scanlon’s account of *blame* (2008: pp. 128–129). This paper investigates what we can learn from the approaches to conceptual re-engineering that were developed by Carnap and Goodman, which are known respectively as explication and reflective equilibrium.

Explication is enjoying a revival. Among Carnap scholars, the discussion has been propelled by Carus’s (2007) thesis that explication is the key to a promising philosophical programme in the tradition of the Enlightenment. And the view that explication is a prominent method in philosophy has gained currency and sparked various projects of explication (see, e.g. Baumberger forthcoming; Cappelen forthcoming; Dutilh Novaes and Reck 2017; Haas 2015; Kuipers 2007; Leitgeb 2017; Olsson 2015; Pinder 2017; Shepherd and Justus 2015; Vermeulen 2013). However, the discussion has been dominated by ‘external’ challenges, that is, arguments which call into question the value of explication as a philosophical method, and specifically the worry, first voiced by Strawson (1963), that explications are philosophically unhelpful because they aim at replacing concepts, and hence change the subject instead of clarifying or analysing *our* concepts (e.g. Maher 2007; Justus 2012; Schupbach 2017). This paper does not aim at contributing to this debate, which I think suffers from two shortcomings: it too often relies on sketchy descriptions of explication, and it is too narrowly focused on the subject-change challenge. As a result, the merits of the method are difficult to assess and internal problems of available accounts of explication have been overlooked. To make progress, we need to go beyond Carnap’s classic exposition. As I will show, some key resources can be found in Goodman’s work on reflective equilibrium.

Although reflective equilibrium is often considered a standard method of philosophy (see Daniels 2016; Hahn 2000; Stein 1996), the debate about reflective equilibrium similarly suffers from relying on sketchy accounts.[2] As I will argue, not only is reflective equilibrium a further development of explication, but a great deal can be learned from the systematic connections between reflective equilibrium and explication. To

[1] I use italics to indicate that “*terrorist*” refers to the concept of being a terrorist, not to terrorists.

[2] Ironically, the most elaborate accounts, namely Elgin’s (1996) and Tersman’s (1993), are the least discussed ones.

my knowledge, these connections have never been discussed in detail. Occasionally, a link between Carnap's and Goodman's methodological ideas is at least mentioned (e.g. Cohnitz and Rossberg 2006: p. 62; Lutz 2012: p. 192). But the textual evidence that links Goodman's discussion of reflective equilibrium with Carnap's method of explication seems to be virtually unknown. Sometimes it is even argued or simply presupposed (e.g. Dutilh Novaes and Geerdink 2017; Miller 2000; Nagel 2007) that explication and reflective equilibrium are fundamentally distinct. One explanation might be that it is well known that reflective equilibrium is a method of justification, but not that Goodman tied reflective equilibrium extremely close to conceptual engineering and rational reconstruction.

There are further reasons why the link from Goodman's discussion of reflective equilibrium to Carnap's account of explication is easily overlooked. One is that this link runs indirectly through Goodman's lesser known theory of constructional definition, which is the methodological account he developed in *The Structure of Appearance* (1977; henceforth "SA") in a direct reaction to Carnap. In *Fact, Fiction, and Forecast* (1983; henceforth "FFF"), Goodman then declares that his remarks on reflective equilibrium elaborate on certain aspects of his theory of constructional definition. However, what Goodman says about reflective equilibrium not only builds on his theory of constructional definition, but also involves shifts in perspective and terminology. An adequate understanding of Goodman's methodological thinking must therefore draw on an analysis of his account of constructional definition, his discussion of reflective equilibrium and the relation between the two.

Finally, one might think of explication, constructional definition and reflective equilibrium as methods that were developed in the context of different projects and hence as methods serving different purposes. Carnap worked on theories of logical probability and semantics, and Goodman on constitutional systems and the justification of inductive logic. Nonetheless, they both saw their methodological proposals as contributions to a method of conceptual re-engineering that is not tied to their specific projects, but broadly applicable in philosophy and science.[3]

Taking this point seriously has strategic consequences for this paper. It is a reason for focusing on Carnap's and Goodman's explicit methodological considerations rather than trying to distil methodologies from what they actually did in their specific projects. The latter would run the danger of building specific features of these projects into the methodology although Carnap and Goodman did not conceive of them as aspects of their methodology. Carnap, for example, was working with formal methods, but did not consider this to be an essential aspect of explication (Carnap 1963b: pp. 936, 937), and in SA Goodman went to lengths to develop nominalistically acceptable concepts although he earlier (1990) used the same method without nominalist ambitions. Of course, it would be interesting for independent reasons to have case studies which relate my analysis and the methodological proposals I develop here to what Carnap, Goodman and other philosophers have actually done. But such studies must clearly be left to further papers.

[3] See the remarks in Carnap 1963b: pp. 933–939; SA:ch. I.1–2; FFF:66–7.

In what follows, I investigate the structural interconnections between Carnap's and Goodman's methodological reflections and explore the systematic potential of combining them. One result will be that the historical and systematic links are so close that the various methodological proposals are best interpreted as contributions to an overarching philosophical project, each elucidating certain aspects of conceptual re-engineering for theoretical purposes. More importantly, the structural links give us reason to re-think the standard accounts both of explication and reflective equilibrium. In this way, the ideas of Carnap and Goodman enrich each other and we can draw a number of consequences for an improved account of conceptual re-engineering that is promising as a method of philosophy. Three main points will be argued for. (i) An adequate account of a method of conceptual re-engineering for theoretical purposes must focus not on individual concepts but on systems of concepts and theories; this becomes most clear in Goodman theory of constructional definition. (ii) The linear structure Carnap describes must be replaced by a process of mutual adjustments as explained by Goodman. (iii) Goodman's description of reflective equilibrium must be supplemented by criteria of adequacy which are prominent in the method of explication and which ensure that the resulting concepts promote theoretical virtues while being sufficiently similar to the concepts we started out with. Being driven by these two typically antagonistic aspirations—similarity and theoretical usefulness—emerges as the heart of conceptual re-engineering for theoretical purposes. Eventually, the results (i)–(iii) place conceptual re-engineering firmly into the process of theory development.

Section 2 briefly revisits the history of explication and reflective equilibrium, and presents some textual and historical evidence for the links between Carnap's and Goodman's methodological investigations. The three following sections analyse the structure of their methods and the criteria of adequacy they propose. Some exegetical work is needed to expose the systematic points that shed new light on explication and reflective equilibrium. I begin with an exposition of Carnap's method of explication and present an interpretation that gives priority to his later, more pragmatic explanations (Sect. 3). Goodman's ideas are then introduced as answers to shortcomings in Carnap's account of explication. Section 4 discusses Goodman's theory of constructional definition, in which he attacks Carnap's similarity requirement for explications. My analysis shows that Goodman in effect also introduces structural refinements into the theory of explication. Section 5 turns to Goodman's discussion of reflective equilibrium, in which he explicitly addresses the non-linear structure of explication and constructional definition. However, linking reflective equilibrium to the theories of constructional definition and explication requires some analytical efforts (Sect. 5.3). It also becomes clear that FFF gives an incomplete characterization of reflective equilibrium, which must be read against its background in Goodman's and Carnap's theories of definition and explication (Sect. 5.4). Finally, Sect. 6 sketches a research agenda for reflective equilibrium as a method of conceptual re-engineering for theoretical purposes.

2 Historical and textual links between reflective equilibrium and explication

Reflective equilibrium and explication are closely related historically. This becomes apparent when we ask why Rawls (in 1999c) rather than Goodman coined the label "reflective equilibrium" ("RE" for short), even though Goodman's classic description of RE in FFF pre-dates *A Theory of Justice* by almost twenty years.[4] The answer is that Goodman simply had no need for a new term. When he discussed RE as the method for justifying deduction and induction, he spoke of "definition",[5] and did so for clear reasons:

> The task of formulating rules that define the difference between valid and invalid inductive inferences is much like the task of defining any term with an established usage. [...] Thus the interplay we observed between rules of induction and particular inductive inferences is simply an instance of this characteristic dual adjustment between definition and usage, whereby the usage informs the definition, which in turn guides extension of the usage. (FFF 66)

As this quotation shows, Goodman thought of his remarks in FFF as elaborating an aspect of the theory of definition; specifically, of his theory in SA, as he explicitly says a few lines further down (FFF 67n3).[6]

Goodman's theory of definitions in SA, in turn, is clearly associated with Carnap's programme of explication and rational reconstruction. After all, Goodman originally developed his theory of definitions in *A Study of Qualities* (1990), in the context of his critical discussion of Carnap's programme of reconstruction in the *Aufbau* (2003). That Goodman's programme was also related to Carnap's later conception of explication from *Logical Foundations of Probability* (1962; henceforth "LFP")[7] is not confirmed by explicit statements in SA or FFF, but I surmise that this is primarily because at the time it was just evident to Goodman, Carnap and their interlocutors. In fact, there is considerable evidence for the close association of Goodman's account of RE and his

[4] The following analysis covers only part of the development that led to Rawls's exposition of the method of RE in *A Theory of Justice.* Rawls built on at least four sources: the logical empiricists' work on epistemology, inductive logic and physicalism (see Rawls 1950:ch. I.6); his own earlier work (1950, 1999b, 1999d); Chomsky's conception of competence and performance (1965:§§ 1–2); and Goodman's views published in *Fact, Fiction, and Forecast* (mentioned in Rawls 1999c: p. 18n7). There are studies on the development of Rawls's methodological thinking (e.g. Mikhail 2010; Mäkinen and Kakkuri-Knuuttila 2013) and on Chomsky's influence (e.g. Daniels 1980; Mikhail 2010), but although it is well known that Rawls's ideas are related to Goodman's, this historical line has never been studied in detail.

[5] Goodman does not use "reflective equilibrium", but "equilibrium" can be found in 1968: p. 163. Quine (1980: p. 43) and Scheffler (1954: p. 187) had already used "equilibrium" and "reequilibrating" with respect to justification.

[6] Astonishingly, this explicit link from RE to Goodman's theory of definitions in SA has virtually gone unnoticed in the literature. So it may be worth pointing out that Goodman's exposition of RE in FFF is not limited to ch. III.2, which contains the invariably quoted passages on mutual adjustment and virtuous circularity: it is continued in the opening paragraphs of ch. III.3, and there Goodman links RE to his theory of definitions.

[7] §§ 2–3 of LFP are Carnap's classic exposition; an earlier version is 1956: § 2; the term "explication" first appears in 1945.

theory of definitions with Carnap's method of explication. Firstly, there are terminological clues. When Goodman speaks of his own theory in SA, he sometimes uses "explicative definition" as a stylistic variant for what he usually calls "constructional definition", "logical construction" or simply "definition".[8] And in FFF, he incidentally describes the definitions sought by the method of RE as "explications" (FFF 65n2; also 46–8). Secondly, Goodman's contribution to *The Philosophy of Rudolf Carnap* (Goodman 1963: pp. 555–556) refers to Carnap's reconstructions in the *Aufbau* as "constructional definitions", and Carnap in his *Replies and Systematic Expositions* (1963b; henceforth "RSE“) explicitly identifies Goodman's definitions with explications (p. 945). Finally, the link between RE, Goodman's theory of definition and Carnap's explications was noted almost immediately after their publication by Hempel (1953; see also 2000: pp. 206–208).[9]

This evidence shows that Carnap's method of explication, Goodman's theory of definitions and his account of RE should be read as contributions to an overarching project of developing a philosophical account of conceptual re-engineering and, more generally, theory development.[10] Sections 3– 5 now examine Carnap's and Goodman's methodological proposals in detail.

3 Carnap's method of explication

3.1 The basic ideas

In a nutshell, Carnap describes explication as a process which replaces, for some theoretical purpose, an inexact concept (the explicandum) with a more exact concept (the explicatum) which is explicitly introduced into the system of concepts of a target theory.

One of Carnap's examples is the explication of the concept *fish*, which is suitable for everyday purposes but less so for biological theory, where it was replaced by the concept *piscis*, characterized in biological terms as "cold-blooded aquatic vertebrate" (LFP §3).[11] Another example is the quantitative concept of temperature, which replaces the everyday concepts *cold*, *warm*, *hot* etc. for scientific purposes (LFP §§4–5). Indeed, countless explications can be found in science and philosophy; the examples mentioned in the Introduction are only the tip of the iceberg.

[8] e.g. SA:L, 19n8; Goodman 1972a: p. 4. Ch. I of SA is titled "constructional definition" but indexed as "constructional or explicative definition".

[9] Since then, the connection has occasionally been recognized in the literature (e.g. Hanna 1968: p. 29; Hellman 1977: pp. XXVII–XXVIII; Kantorovich 1993: p. 122; Elgin 1997; Cohnitz and Rossberg 2006: p. 62), and an explicit link from RE to (Quine's account of) the method of explication can be found in Rawls (1999c: pp. 95–96). But to my knowledge no in-depth analysis has been given so far.

[10] Since this section aims only at substantiating the thesis of a common project, other historical issues must be left unaddressed, specifically the sources on which Carnap and Goodman built (e.g. Russell's (1993; 1954:ch VI–VIII) and Whitehead's (1919) logical constructions), the relation of Carnap's and Goodman's proposals to related ideas put forward by contemporaries (e.g. Hempel 1952; Pap 1949; Quine 1960) and questions of precedence, for example, who first came up with the example of fish and whale in this context (Goodman used it in 1990: pp. 49–50).

[11] Of course, this example is tailored to expository purposes, not a piece of up-to-date biological taxonomy.

Explicating is a two-step process. The first is called "clarification of the explicandum" and seeks to identify the explicandum as clearly as possible. If the explicandum-term is ambiguous, it must be disambiguated. Since an inexact explicandum cannot be defined exactly, it must be characterized informally, for example, by specifying cases in which the explicandum clearly does or does not apply. The goal is to avoid arguing at cross-purposes by making sure that it is at least practically clear how the explicandum is used in ordinary cases in the relevant contexts (LFP 4). Secondly, an explicatum must be introduced by specifying explicit rules for using the explicatum in terms of the target system of concepts. This can be done by defining the explicatum, or by some other method of concept introduction. Since the explicandum is inexact, there is in general not just one correct explicatum, but several more or less adequate explicata can be given for the same explicandum. Their adequacy must be assessed in light of the role the explicatum is expected to play in the target theory and involves several criteria. Carnap names four: similarity to the explicandum, exactness, fruitfulness and simplicity. They are all a matter of degree and potentially subject to trade-offs.

3.2 The structure of explications

With respect to the structure of explications, we need to clarify three points that have been debated in the literature.[12] Firstly, explications involve two systems of concepts. Whereas the explicandum is a pre-theoretical concept in the sense of belonging to everyday language or to a preceding stage of theory, the explicatum belongs to the system of concepts which is tied to the scientific or philosophical theory in which the explicatum is going to be used. This does not imply that a formal language is used for the target system of concepts, as the example of *piscis* shows.

Secondly, there is some disagreement on whether explication deals with terms, concepts or both (see, e.g. Maher 2010). This unclarity arises because "explicandum" is sometimes used to refer to the input of the first step of explication, the clarification of the explicandum, sometimes to its output and sometimes to both. Since the input is possibly ambiguous, it must be a term, given that ambiguity is standardly understood as arising if a term has more than one meaning. And since the output of a successful clarification must be unambiguous, it cannot be simply a term. A plausible description of the step of clarification is therefore that it takes as its input a term, the "explicandum-term", and identifies a concept, the "explicandum", by identifying one specific way of using the explicandum-term.[13] Consequently, the explicatum must be a concept as well since it is meant to replace the explicandum. The resulting structure of explications is depicted in Fig. 1.

[12] See Brun 2016 for a more extensive discussion of many of the points made in the rest of this section.

[13] Carnap identified concepts with certain non-linguistic abstract entities (LFP 7–8) and thereby tied his account of explication to his semantics. Since nothing in his account depends on whether we adopt his view of concepts, I generalize it by using "concept" to refer to an elementary linguistic entity, a "term", together with rules for its use. And I take a neutral stance on the nature of such rules; they may specify a term's intension or extension; they may be stated explicitly or be given implicitly in usage, fairly clearly or rather turbidly. I also leave open how such rules are related to mental or abstract entities which are often called "concepts".

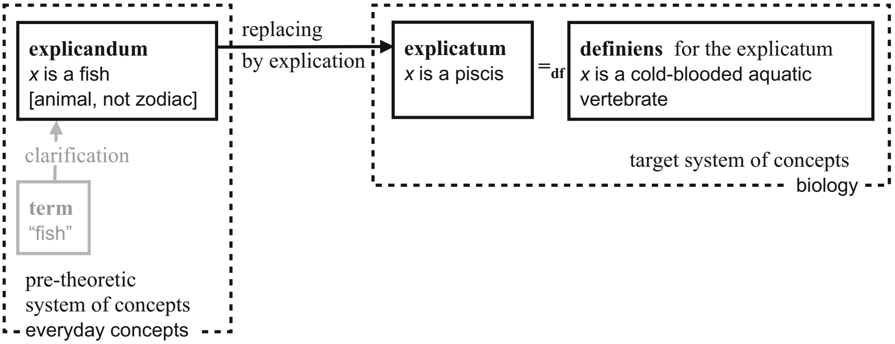

Fig. 1 Explication of *fish* by means of a definition (adapted from Brun 2016)

This diagram, thirdly, also shows how explications relate to definitions. Although explications can include a definition, they are not definitions (contra, e.g. Cohnitz and Rossberg 2006: p. 58). In the example, the explicatum is the definiendum, and the definiens says how the explicatum-term "*x*is a piscis" is to be used. This definition is given within the target system of concepts and does not include the explicandum. Hence, it is inappropriate to think of an explication as defining the explicandum.

3.3 The criteria of adequacy

Carnap's four criteria of adequacy also call for further explanations. To begin with, some differentiations and additions are needed with respect to exactness, fruitfulness and simplicity. Carnap's use of "exact" covers several qualities of an explication. Some of them are necessary conditions for the adequacy of any explication, namely that the explicatum is consistent and that explicit and unambiguous rules for using the explicatum-term are given in terms of the target system. Consistency ensures that no paradoxes or contradictions occur as a result of the rules for using the explicatum-term dictating that the explicatum both applies and does not apply to the same objects. Exactness in a narrower sense is the opposite of vagueness, meaning that the rules should allow in as many cases as possible a clear decision on whether the explicatum does or does not apply.

By "fruitfulness", Carnap means that it should be possible to formulate many laws or other generalizations featuring the explicatum, and by "simplicity" that the rules for using the explicatum and the laws which include the explicatum be simple. Clearly, these are just two examples of theoretical virtues we may expect an explication to promote, and Carnap (LFP §5) in fact refers to further virtues such as the wider scope of the resulting theory, and precision in the sense of more precise descriptions and finer discriminations being made possible by quantitative concepts. Other important virtues are the explanatory power of a theory, and its ability to be used for predicting novel phenomena or for deciding on practical problems.[14] In addition to such generally rel-

[14] Later on, especially after Kuhn (1977), desiderata for theories became the focus of extended discussion (see Douglas 2013 for a short survey).

evant desiderata, explicata are typically also expected to fulfil some specific functions in the target theory, which then also count as criteria of adequacy. In *Meaning and Necessity* (1956: pp. 8–12), for example, Carnap seeks an explication of *necessary truth* that can be used as a basis for explicating *entails*, *is factually true* and a range of further semantic concepts.

Furthermore, the criterion of similarity raises two questions. Is it to be understood in terms of extensions? Is it the same for all explications? And similar issues arise for exactness. At this point, it is important to note that there are tensions and developments in Carnap's writings with respect to similarity and exactness. In LFP, Carnap tends to explain similarity in terms of the extensions of explicandum and explicatum, and he emphasizes that explications should reduce vagueness (LFP 5). Specifically, he held that similarity requires at least overlapping extensions, although differences in extension are permitted. On this basis, many interpreters claim that reducing vagueness while preserving clear-cut cases is a core idea of explication (e.g. Hanna 1968; Hempel 2000). However, Carnap leaves room for alternative explicata with disjoint extensions (e.g. Zermelo's and von Neumann's explications of the number *two*; see Carnap 1956: p. 8), and he prominently refers to cases in which a more fruitful explicatum is obtained not by reducing vagueness, but by reclassifying clear-cut cases (*piscis* excludes whales and dolphins). Hence, neither overlapping extensions nor reducing vagueness is indispensable for adequate explications.

Carnap's later explanations in RSE are more pragmatic in two senses. He emphasizes that explicating requires practical decisions in light of the problems the new concept is expected to solve and the role it is supposed to play in the target theory. Moreover, the criteria of adequacy are dependent on the target theory and the specific purposes the explication is supposed to serve. Although we can still assume that explications must not increase vagueness, reduction of vagueness is much less prominent in RSE, and Carnap's insistence on "exactness" first of all means that the explicatum is unambiguous, consistent and introduced by explicit rules. He also makes clear that similarity basically requires that the explicatum can be used in place of the explicandum in relevant contexts. This leaves room for interpreting the similarity criterion differently in different situations, depending on the purpose the explicatum is supposed to serve. Only for specific projects of conceptual re-engineering can we decide on the relative weight of the various theoretical virtues, and on the price they are worth paying in terms of similarity. Elsewhere (Brun 2016) I argue for giving priority to Carnap's later view and using it as a guide for interpreting LFP, but for present purposes we need to put on record that Carnap's writings, especially LFP, also provide reason to find the exact nature of his similarity criterion unclear.

3.4 Two limitations of Carnap's account

Carnap's account of explication has two limitations which are relevant in the present context.[15] Firstly, it focuses on replacing individual concepts, although there are several reasons why explications must be dealt with in a more comprehensive setting.

[15] A third limitation is discussed in Brun (2016).

Further concepts are needed to characterize an explicatum. If they are to be introduced into the target system by explication as well, this calls for orchestrating explications. And in many projects of explication the target theory is not readily available, and explicating concepts must therefore go hand in hand with developing a target theory. Finally, many aspects of adequacy are chiefly related to systems of concepts and theories. Examples include not only consistency, scope of application, explanatory power and further virtues of theories, but also Carnap's fruitfulness and simplicity. These points show that explications must ultimately be understood as "building blocks" for theory development. And this is, of course, how explications are used in Carnap's own work on probability, logic and meaning. Nonetheless, he did not give an explicit account of a method for explicating entire systems of concepts or theories.[16]

The second limitation concerns the structure of the process of explication, which Carnap basically describes as a linear sequence of steps. However, it is not uncommon that earlier steps of an explication need to be revised in the light of an attempt to introduce an explicatum which shows, for example, that further clarification of the explicandum is needed or that the envisaged criteria of adequacy are not jointly satisfiable. It may even be that the resources of the target system of concepts are needed to complete the first step of an explication, for example because a subtle ambiguity of the explicandum can only be detected with the help of exactly defined alternative explicata. Moreover, every successful explication includes a "feedback" effect. For those purposes for which the explication is undertaken, the use of the explicandum is discontinued in favour of the explicatum. In some cases this means that the explicandum is no longer used at all. A linear structure as depicted in Fig. 1 fails to make these aspects of explication transparent.

One might object that Carnap was well aware that the practice of explicating may call for proceeding in a non-linear way. This is plausible. At one place, he (Bar-Hillel and Carnap 1953: p. 150) discusses a case in which some proposed criteria of adequacy need to be revised because it turns out that they cannot be satisfied simultaneously. At another, he briefly mentions that explications can have the side-effect that the explicandum-term changes its meaning in everyday usage (e.g. *x is warmer than y* is understood in the sense of its scientific explicatum *the temperature of x is higher than that of y*; LFP 12–3). In spite of such remarks and allusions, Carnap's explicit methodology describes only a linear procedure. A similar point can be made with respect to the view, recently made prominent by Carus (2007:xi, 19–21), that Carnap was in fact aiming at a "dialectical", non-linear, method (see also Reck 2012). Since Carus himself says that Carnap left this dialectics implicit, Carus's view is compatible with my analysis of Carnap's explicit methodology.

The following two sections analyse how Goodman's work can be interpreted as picking up on the problems pointed out in this and the preceding subsection. In his theory of constructional definition, Goodman attacks the idea of explaining similarity with reference to overlapping extensions, and proposes an alternative criterion which

[16] Actually, many authors are ready to speak of the explication of, for example, systems of concepts or theories. But explicit discussions of how Carnap's ideas could be extended to more complex explicanda have been scarce. Exceptions include Hegselmann (1985) and Martin (1973), but they assume that reducing vagueness is *the* goal of explication.

is tailored to deal not with individual concepts but with systems of concepts. And in his work on RE, he develops an alternative to the linear structure of the process of explication. Although Goodman considered what he said on RE as a contribution to the theory of constructional definition, linking the relevant parts of SA and FFF requires to work through considerable shifts in perspective and terminology. I therefore discuss Goodman's methodological proposals in SA and FFF independently before I analyse how they can be integrated into one account.

4 Goodman's theory of constructional definition and the similarity requirement

4.1 Refining the structure of explications

Goodman opens his discussion of conceptual re-engineering (in ch. I of SA, on "constructional definition") with an attack on similarity criteria that require overlapping or identical extensions.[17] He argues that explicandum and explicatum can have disjoint extensions: in geometry, for example, points may be defined as pairs of intersecting lines or as certain sets of volumes, but points are not pairs of lines, nor are they sets of volumes (SA 5–7). Goodman concludes that similarity must be judged by a weaker criterion and suggests a standard he calls "extensional isomorphism".[18] The basic idea is that similarity requires only that explications preserve extensional properties of and relations between explicanda, not necessarily their extensions. For the moment, further details of Goodman's criterion do not matter (I will say more about it in the next subsection), except that it implies that the extensions of the explicandum and the explicatum have the same cardinality; that is, that their extensions have the same number of elements. For example, every point corresponds to exactly one pair of intersecting lines and vice versa.

This description suggests that Goodman's constructional definitions have the same structure as Carnap's explications, but weaker criteria of similarity. Yet even though this is how Goodman and Carnap frame the issue (in RSE 945 and Goodman 1963: pp. 554–556), it is not quite accurate. The problem becomes apparent if we ask which relation should be subject to Goodman's criterion of extensional isomorphism. No plausible candidate can be found in Fig. 1. For the definition of the explicatum within the target system of concepts, it makes no sense to require only extensional isomorphism, because this very definition establishes extensional equivalence.[19] And for the

[17] Some remarks on terminology: following Goodman, I refer to his method of conceptual re-engineering as "constructional definition". But I avoid his use of "definition", "definiens" and "definiendum" and continue to speak of "explication", "explicandum" and "explicatum" (also because, as will shortly become clear, constructional 'definitions', just like explications, have a more complex structure than definitions in the usual sense). I also continue to use Carnap's "similarity" despite Goodman's attacks on this notion (see 1972b). In the present context the notion of similarity does no substantial work; it has to be spelled out in terms of the explicandum's features an explication is supposed to preserve.

[18] Goodman does not claim that extensional isomorphism is the criterion of similarity for every kind of explication, but leaves open the possibility that in some cases other criteria are needed (1972a: p. 84).

[19] We can assume that the explicatum is introduced by a definition, since Goodman deals with this case only.

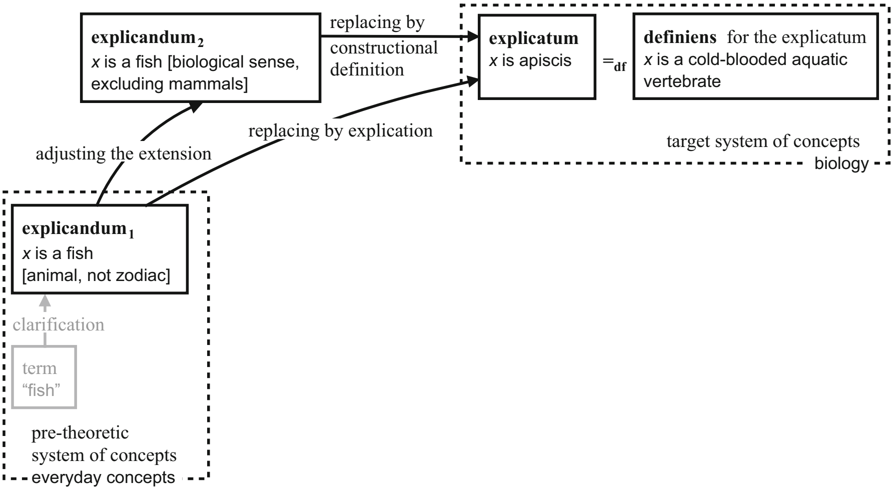

Fig. 2 Constructional definition and explication of *fish*. (I retain the preliminary step of clarification from Carnap's account of explication although Goodman does not clearly distinguish the two steps *clarification* and *adjusting the extension*)

relation between explicandum and explicatum, extensional isomorphism (or extensional equivalence, for that matter) cannot be required since explications often aim at solving problems by adjusting the extension of a concept in a way that changes the extension's cardinality: there are more fish than pisces.[20]

The problem can be solved if we recognize that Goodman's proposal in fact requires a modification to the structure of explications described in Sect. 3. If we insert an additional element—let's call it "explicandum$_2$"—between explicandum ("explicandum$_1$") and explicatum, we can distinguish between two tasks which are performed in an explication. First, the extension of the explicandum$_1$ is adjusted, then the resulting explicandum$_2$ is replaced by the explicatum (see Fig. 2). For the latter task extensional isomorphism is a plausible requirement.

That Goodman's account in fact relies on such a refined structure is not obvious in his description, because he uses "definiendum" in two different ways. At the beginning of SA (pp. 4–5), when he briefly deals with vagueness, ambiguity and the "trimming and patching" (i.e. adjusting the extension) of concepts for scientific use, he uses "definiendum" to refer to the explicandum$_1$. As soon as he turns to his main subject, extensional isomorphism, "definiendum" refers to the explicandum$_2$. The change is not made explicit, and instead of sorting out the matter, Goodman rather glosses over the lack of clarity in his text. Only in a footnote, inserted into later editions (SA 19n8), he remarks that his reference to "presystematic" concepts is supposed to be understood as reference to concepts whose extensions have already undergone adjustment; that is, as reference to the explicanda$_2$ in the terminology of the present paper.[21]. This remark

[20] A similar argument can be made in the case of explications that reduce vagueness.

[21] The footnote explains "presystematically" as "according to the understood or express informal explication of what is to be defined. This explication normally accords in general with ordinary usage, trimmed and

of Goodman's also applies to his actual constructive work in the latter parts of SA, where his reference to "presystematic" use of terms refers to explicanda$_2$, not simply to everyday concepts.

Figure 2 also shows how Carnap's and Goodman's methods of conceptual re-engineering can be related to each other. Both replace a concept for theoretical purposes, but from a Goodmanian point of view, Carnapian explication deals with two tasks (adjusting and replacing) simultaneously, whereas from a Carnapian perspective, Goodman splits the relation between explicandum and explicatum in two parts. Consequently, where Carnap has one criterion, Goodman needs to distinguish two aspects of similarity.

4.2 Two aspects of similarity

The distinction between two tasks and two criteria of similarity can now be used to shed new light on conceptual re-engineering. When Goodman very briefly deals with the first task, adjusting the explicandum, he mentions ambiguity, vagueness and inappropriateness for scientific use as three types of problems that need to be tackled before the isomorphism criterion can be applied (SA 4–5).[22] In Carnap's account, they correspond to the demands for unambiguous, exact and fruitful concepts. Whereas ambiguity is dealt with in the clarification of the explicandum, the latter two provide reason to diverge from the extension of the original explicandum. Reducing vagueness calls for settling borderline cases, and replacing concepts by theoretically more fruitful ones often requires adjusting extensions. Such adjustments raise questions of similarity. The extension of a concept may be adjusted for the sake of theory, but it may not be changed to any extent whatsoever. We clearly expect the extension of the adjusted explicandum$_2$ to overlap with the extension of the original explicandum$_1$.[23] After all, adjusting the extension of the explicandum must not amount to simply changing the subject.[24] It seems, however, illusory to ask for one precise yet generally applicable

Footnote21 continued
patched in various ways for various purposes" (SA 19n8). This explanation was introduced in the second edition (SA 1966, 25). In the first edition (SA 1951, 22), it simply reads "I use 'presystematically' for 'according to ordinary usage"', which is not quite correct as we have seen. Goodman's use of "informal explication" may evoke misunderstandings, such as that the concept he calls the "definiendum" may be the product of clarifying the explicandum or of an explication in Carnap's sense. Goodman also does not use "presystematic" is the sense in which I use "pre-theoretical". "Presystematic" does not imply that the explicandum$_2$ is in actual pre-theoretical use; it merely emphasizes that the explicandum$_2$ is not part of the target system of concepts.

[22] I will often use "adjusting the explicandum" as an abbreviation for "adjusting the extension of the explicandum". "Adjusting" is meant to include the limiting case of leaving the extension of the explicandum unchanged.

[23] See Goodman's remark ("normally accords in general") quoted in note 22.

[24] One might see this as addressing Strawson (1963) subject-change challenge, but this is not intended (I discuss Strawson's challenge in Brun (2016)). The requirement of extensional overlap gives the worry of a "subject change" a related but different interpretation. For Strawson, the explicandum is the subject and hence replacing it by another concept always amounts to changing the subject. Constructional re-engineering, on the other hand, aims at replacing concepts and the condition in the main text only expresses that a complete change in extension would result in changing the subject.

criterion of similarity in this sense, also because each explication will have to strike a balance between similarity and the other aspects of adequacy. As a general instruction, Carnap's statement that similarity requires that the explicatum can be used in place of the explicandum in relevant contexts remains appropriate. For specific explications, we should, of course, try to specify conditions of similarity explicitly, for example by specifying relations to other concepts or by giving clear positive and negative instances to which we expect the explicatum (not) to apply.[25]

The second task consists in replacing the adjusted explicandum$_2$ by an explicatum. The criterion of similarity for this step is Goodman's main concern in chapter I of SA. The standard he defends, extensional isomorphism, means that for a system of explications there must be an injective, but not necessarily surjective, function from the set of objects that make up the extensions of the explicanda$_2$ to the set of the elements of the extensions of the explicata (SA 10–1).[26] Roughly speaking, this criterion places no restriction on what objects—primitive or defined in the target system—substitute for the objects in the extension of the explicanda$_2$, but it ensures that all extensional properties (e.g. reflexivity or symmetry) of and relations between the explicanda$_2$ (e.g. inclusion) also hold for their counterparts in the target system of concepts. Extensional isomorphism is therefore a source of pluralism since it permits various extensionally non-equivalent explicata for the same explicandum (Goodman 1963: pp. 555–556).[27]

To illustrate these points, we can pick up Goodman's geometrical examples. The criterion of extensional isomorphism permits that the concept *point* be explicated in a number of different ways, for example, by concepts that refer to pairs of intersecting lines, to classes of volumes, or to convergent series of spheres. Relations such as *x and y are coincident* and *x is to the left of y* can also be explicated in many ways. But since extensional isomorphism demands that extensional properties and relations be preserved, explicata for *x and y are coincident* must be reflexive, symmetrical and transitive, and explicata for *x is to the left of y* must be converses of explicata for *x is to the right of y.*

As Goodman emphasizes, it is important to note that the criterion of extensional isomorphism is to be applied to an entire system of explications simultaneously, not to each explication individually (SA 16; Goodman 1963: p. 556). In fact, extensional isomorphism is a plausible criterion only if understood in this way. Applied to an individual explication, it merely demands that the extensions of the explicatum and the explicandum$_2$ have the same cardinality and structural properties. But the criterion quickly gets more demanding when a system is established by a number of interlinked explications. In Goodman's example, we can imagine starting with an explication of *point* in terms of intersecting lines; this will have consequences for a subsequent

[25] In passing, we may note that Carnap's fourth criterion, simplicity, is dealt with under the labels "naturalness" and "technical efficiency" in SA (not to be conflated with simplicity as analysed in ch. III of SA). They refer to the psychological ease of handling the explicata and the degree of complexity that is needed for introducing all the explicata we want to introduce (SA 19–20).

[26] Goodman does not use "isomorphism" in its usual sense, which implies a bijection. For discussions of Goodman's notion of extensional isomorphism, see Hellman (1977: pp. XXVII–XXXI, 1978).

[27] Note that this form of pluralism differs from pluralism resulting from different ways of adjusting the extension of the explicandum.

explication of *to the left of*, the extension of which will now need to include certain pairs of pairs of intersecting lines; and this explication will in turn restrict our choice of explicata for *to the right of* to converses of the relation *to the left of*; and so on for further explications.[28]

We can now relate the analysis of constructional definitions in the preceding paragraphs to Carnap's account of explication. Firstly, Goodman shifts attention from individual concepts to systems of concepts by introducing a similarity criterion which is explicitly tailored to systems of concepts. He thereby makes a crucial step to overcome a limitation of Carnapian accounts of explication, which address only individual concepts explicitly. This shift is an improvement upon Carnap's explicit methodology, which is very much in line with Carnap's actual work. Carnap developed theories of probability and semantics, not individual explications, and he sometimes refers in his examples to entire sets of explications (e.g. LFP 15–18). But only Goodman drew the explicit methodological consequence of focusing on systems of concepts.

Secondly, we can now offer an explanation of why Carnap had divergent interpretations of the criterion of similarity. In Sect. 3.3 we found that in his earlier work in LFP, Carnap was inclined to think in terms of extensional overlap or identity with respect to clear-cut cases. But later on, he took a more pragmatic stance and held that different criteria of similarity may be appropriate in different contexts. Whereas extensional isomorphism may be the right standard in some situations, stronger criteria such as extensional equivalence or even synonymy may be more appropriate in others (RSE 945). Against the backdrop of my analysis of constructional definitions, these two divergent views can be explained by a shift in perspective. Whereas in LFP, Carnap was more concerned with the first task of adjusting the explicandum, in RSE, he focuses more on the second task of replacing the adjusted explicandum by an explicatum. As I have argued above, the former calls for similarity in the sense of extensional overlap, whereas the latter leaves room for a criterion such as extensional isomorphism.

4.3 Two limitations of the theory of constructional definition

As we have seen, extensional isomorphism deals only with part of what Carnap's similarity criterion for explication covers. It presupposes that ambiguity and vagueness have been dealt with and that any other trimming and patching of extensions that might be appropriate for theoretical purposes has been done as well (Hempel 1953; see Cohnitz and Rossberg 2006: p. 253n26). In practice, however, the extension of the explicandum$_2$ usually cannot be specified in an exact manner in advance. This does not invalidate extensional isomorphism as a criterion of similarity, but rather means that in practice the two steps of adjusting and replacing the explicandum may not be neatly separable. Usually, the explicandum$_2$ cannot be specified independently, but only indirectly by introducing an explicatum. We simply give a definiens for the explicatum

[28] For an extended discussion of Goodman's example, see SA 10–16. More examples can be found especially in Part Two of SA, where Goodman uses his methodology of constructional definition to actually develop a constitutional system as an alternative to Carnap's *Aufbau* (2003).

and thereby commit ourselves with respect to the extension of the explicandum$_2$. Judging whether the proposed adjustment of the explicandum is acceptable will then not be possible before an explicatum is tentatively introduced.

This observation draws attention to two respects in which Goodman's theory of constructional definition remains unsatisfactory. Firstly, although Goodman makes clear that the "trimming and patching" that leads to the explicandum$_2$ is guided by theoretical virtues (specifically exactness and fruitfulness), he offers virtually no explanation of how such adjustments are in fact effected. And doing so would force him, secondly, to take the non-linear structure of conceptual re-engineering more seriously.

There are two additional reasons why constructional definition cannot be adequately described as a linear process (as shown in Fig. 2). Insisting on specifying the extension of the explicandum$_2$ as exactly as possible before proceeding to developing an explicatum would work against the spirit of Goodman's method. It would force us to replace the explicandum$_1$ with an exactly defined concept, and thus to deal with many of the problems that were meant to be solved by introducing an explicatum. Furthermore, Goodman emphasizes that extensional isomorphism generates interdependencies between definitions of explicata. Consequently, the explication of a concept may need to be revised in order to lift restrictions it placed on the available options for subsequent explications (SA 11–5).

In sum, Goodman's account of constructional definition in SA includes an important shift in focus from individual concepts to systems of concepts. And my analysis of his theory reveals that we should distinguish two aspects of similarity. But Goodman's work in SA does not deal systematically with the "trimming and patching" of explicanda and the non-linear structure of conceptual re-engineering. For this, we have to turn to his account of RE.

5 Reflective equilibrium and the non-linear structure of conceptual re-engineering

Goodman presents his idea of a RE in the context of the question of how rules of deductive or inductive inference may be justified (FFF 62–7).[29] A close link to his theory of constructional definition is established when Goodman argues that justifying rules of inference boils down to coming up with an adequate constructional definition of the concept of logical validity. This link may well appear surprising, also because the move from his account of constructional definition to his comments on RE involves shifts in perspective and terminology, which need further explanation. In what follows, I first discuss Goodman's description of RE in FFF and point out some problems that a fully developed account of RE would have to deal with. Sections 5.3 and 5.4 then address its relation to the theory of constructional definition and explication more generally.

[29] Since, according to Goodman, the structure of justification is the same in both cases, I will generally omit "inductive" and "deductive".

5.1 Goodman’s account of reflective equilibrium

At the heart of Goodman’s account of RE in FFF are two key ideas, namely adjustment and agreement:

> I have said that deductive inferences are justified by their conformity to valid general rules, and that general rules are justified by their conformity to valid inferences. [...] The point is that rules and particular inferences alike are justified by being brought into agreement with each other. A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend. The process of justification is the delicate one of making mutual *adjustments* between rules and accepted inferences; and in the *agreement* achieved lies the only justification needed for either. (FFF 64; Goodman’s italics replaced)

The two ideas relate to a process and a state, respectively. As a state, equilibrium is reached when inferences and rules of valid inference *agree* with each other, and this state justifies both inferences and rules of inference. As a process, the idea of a RE refers to the mutual adjustment of particular inferences and rules of inference in a way that leads to the state of equilibrium.

To illustrate Goodman’s ideas, we can look at inferences involving conditionals.[30] Our starting points are uncontroversial examples of valid inferences and rules of inference, such as:

(1) If Aristotle was in Persia, he was in Asia.
Aristotle was in Persia.
Aristotle was in Asia.
(2) Modus ponens: $\varphi \rightarrow \psi; \varphi \vdash \psi$

Since (3*) has true premises and a false conclusion, and is therefore invalid, rule (4*) cannot be accepted:[31]

(3*) If Aristotle lived before Socrates, he lived in antiquity.
Aristotle did not live before Socrates.
Aristotle did not live in antiquity.
(4*) Denying the antecedent: $\varphi \rightarrow \psi; \neg\varphi \vdash \neg\psi$.

Conditionals can raise conflicts. In the context of classical logic, for example, (6) is a rule of valid inference and a conflict arises for those who judge the following inference to be invalid (see Harper 1981):

(5*) If I put sugar in this cup of tea it will taste fine.
If I put sugar and diesel oil in this cup of tea, it will taste fine.

[30] For a more realistic illustration of how the method of RE may be put into action, one could also turn to Goodman’s own work on confirmation theory and induction in FFF. However, although Goodman clearly saw himself as applying RE, he did not explicitly comment on how he worked through subsequent adjustments towards a state of equilibrium. A methodological reconstruction of his actual practice of theory development would require a case study which lies outside the scope of this paper (see Hahn 2000, ch. 2.1 for a relatively short analysis).

[31] Numbers with asterisks are used to refer to inferences or rules which are rejected rather than accepted.

(6) $\varphi \rightarrow \psi \vdash (\varphi \wedge \chi) \rightarrow \psi$

To (re-)establish a state of equilibrium, one can adjust one's stance on (5*) or on (6); that is, accept the inference as valid or reject the rule (a further possibility will be mentioned in Sect. 5.2).

To unpack Goodman's ideas, I first comment on the basic components of RE: inferences, principles, the relation of agreement and the process of adjusting; Sect. 5.2, then discusses some problems in Goodman's account of RE in FFF.

When Goodman speaks about inferences in their relation to principles, he is careful to express that referring to just any inferences will not do when we want to justify principles of valid inference. We rather need to refer to inferences which meet two qualifications.[32] First, there must be a commitment to their validity, expressed either explicitly by a judgement or implicitly in some behaviour such as treating the inference as acceptable in certain kinds of situations.[33] Merely uttering an inference without any commitment concerning its validity (e.g. quoting it as an example of bad style) is irrelevant to the justification of principles of validity.[34] Second, classifying an inference as (in)valid is not to be understood as an isolated act, but as part of a practice, something we normally do and accept when others do it (see Hempel 2000: p. 208). Furthermore, it is important to note that although the discussion in FFF focuses on the concept of *valid inference*, Goodman clearly intends his approach to be as generally applicable to projects of conceptual re-engineering as his theory of constructional definition in SA (as pointed out in Sect. 2). To reflect that in terminology, I speak more generally of "commitments" and "principles" rather than of "inferences" and "rules of inference".[35]

When Goodman speaks about principles, his examples are Mill's "canons" of inductive inference, Aristotle's rules of the syllogism and *Principia Mathematica* (FFF 65–6). These examples show that the relation of agreement is supposed to hold between inferences and a *system* of principles, not just individual principles. On the one hand, the combination of independently acceptable principles may raise problems (e.g. the introduction and elimination rules for "tonk"; see Prior 1960). On the other hand, if we are to assess the agreement between commitments and principles, it is not enough to look at the principles one by one (in this respect, examples (1)–(6) are

[32] When referring to what is justified by being in agreement with logical principles, Goodman simply speaks of "an inference", "a deduction" or "an argument". When referring to what justifies or subverts logical principles, he appeals to "accepted in/deductive practice", "accepted inferences" and "inacceptable inferences", or more explicitly to "the particular inferences we actually make and sanction", "judgments rejecting or accepting particular inferences" and "normally accepted [inductive] judgments" (FFF 63–5).

[33] For example, I can express the same validity commitment by asserting "(1) is a valid inference", and by inferring that Aristotle was in Asia given that I have become convinced that he was in Persia and that he was in Asia if he was in Persia.

[34] This means that Goodman presupposes a pre-theoretical distinction (i.e. a distinction made in everyday language or in a preceding stage of theory) between valid and invalid inferences, but not that he assumes this distinction to be fully embodied in ordinary usage of "valid" or that there is a sharp pre-theoretical distinction between formal validity (i.e. validity in virtue of logical form) and material validity.

[35] I follow Scheffler (1954) and Elgin (1996) in using "commitment" and avoid the more usual "judgement" to block the misunderstanding that commitments must be explicit or conscious. Note that in the technical sense in which I use "commitment" here, commitments involve a propositional attitude of, e.g. accepting or believing, but they need not be strong, acknowledged or reflected.

oversimplifying). Most valid inferences are not instances of just one specific axiom or theorem of, say, *Principia Mathematica*, but rather their validity must be shown by a more complex proof.

The third element, the relation of agreement between commitments and principles, must therefore be understood primarily as a relation between a set of commitments and a system of principles. Although Goodman speaks as if the principles could be tested for agreement one by one, we have just seen that it will often be impossible to pin down a mismatch between commitments and principles to a specific principle and change just this one. Unfortunately, Goodman does not say much about the exact nature of the relation of agreement.[36] But it is reasonable to assume that agreement at least requires consistent commitments, consistency of the principles and that the commitments be inferable from the system of principles.[37] It would also be natural to interpret Goodman as requiring that the system of principles licenses exactly those inferences we are committed to. But this is too strong, because it rules out theories such as zero-order logic (aka propositional or sentential logic) which cover only part of our validity commitments. In fact, giving a detailed account of the notion of agreement raises a range of non-trivial questions, which cannot be addressed in this paper (e.g. how to spell out "inferable", or issues related to logical closure of commitments).

Finally, Goodman describes a process of mutual adjustments: we may revise a principle in light of a strong commitment, or revise a commitment that stands against a principle we are unwilling to change (FFF 64 quoted above), or let the principles decide when we do not have firm commitments (FFF 66; see xviii). This underlines that neither principles nor commitments are immune to revision. Revisions are also not limited to commitments we are unsure of; sometimes "convenience or theoretical utility" may be reason to alter the common usage of a term (FFF 66–7; more on that below). Goodman also emphasizes that the process of adjusting can lead to different systems of principles and commitments in equilibrium (FFF 63).[38] This pluralism results because the process is not specified as a deterministic procedure and involves the creative elements of proposing and adapting principles.

Goodman himself remarks that his description of the process of adjusting simplifies a great deal (FFF 66). Two points are noteworthy. Adjustments should not be thought of as bringing individual principles in line with their particular applications, but more holistically as working towards an agreement between a system of principles and a set of commitments. And when Goodman speaks of our being "unwilling" to alter a commitment or a principle, he assumes that commitments and principles are associated with a certain 'weight', but he does not discuss the nature of such weights (some options are briefly explained in DePaul 2011: p. lxxx).

[36] In FFF 63–5, he speaks only of "conform", "accord", "agree" or "codify".

[37] Consistency may be required for a positive and a negative reason. Agreement is usually interpreted as demanding coherence, which clearly implies consistency. And according to many logics, an inconsistent set of propositions entails every proposition whatsoever, which makes an agreement between inconsistent sets of commitments and principles uninteresting and blocks them from justifying anything.

[38] Goodman emphasizes in many places (beginning with 1990:p. v) that this does not mean that justification is to be had cheaply. It does not imply that we can always justify more than one system of principles and even less that any old system of principles can be justified.

5.2 Three problems of Goodman's account

In FFF, Goodman focuses on principles that explicate just one concept, *logical validity*, but of course we expect a logical theory to explicate further concepts such as *logical truth*, *independence* and *consistency*, and therefore that the method of RE deal with systems of interrelated concepts and principles which effectively constitute a theory. Goodman was certainly aware of this, as not only his choice of examples (e.g. *Principia Mathematica*) confirms, but also his explicit link to systems of constructional definitions in SA and the fact that the construction of systems of concepts and theories takes centre stage in his philosophy. We can therefore safely assume that the exposition in FFF just simplifies when it deals with the justification of logical rules as if that would not only involve but simply consist in explicating the concept of valid inference.

However, this simplification may be partly responsible for a second, more problematic aspect of Goodman's account of RE. Goodman describes the contrast between commitments and principles in terms of an opposition between particular and general.[39] This may seem plausible as long we explicate an individual concept and focus on constructing a concept with a suitable extension. However, there is the fairly obvious objection that commitments can be general as well (Rawls 1999a), for example, the commitment that every inference from a conjunctive sentence to one of its conjuncts is valid. In Sect. 5.3 I will suggest a solution which draws on the link to SA.

Thirdly, the notion of agreement not only raises the problem of partial theories (mentioned above), but also the more fundamental issue of how to handle the relation between commitments and principles, given that commitments may concern ordinary language inferences whereas principles are part of a theoretical system. In the example above, the invalidity of (3*) only counts against (4*) if (3*) is adequately formalized by (4*), and if "⊢" is intended to be interpreted as "entails" rather than, say, "is logically independent of". This may seem obvious, but when deciding whether the invalidity of (5*) counts against (6), one must also address the question whether (5*) can be adequately formalized with material conditionals. Goodman does not systematically deal with such issues, but he draws attention to them in SA (18–9) when he warns against simply identifying the adjusted explicanda$_2$ with their systematic counterparts (the corresponding explicata or the definitions given for them).[40] For present purposes, I will follow Goodman and bracket these issues (see Brun 2014a for a more extensive discussion).

5.3 Linking reflective equilibrium to constructional definition and explication

When Goodman links RE to his theory of constructional definition, he claims that the process of mutual adjustment of commitments and principles described in FFF is "simply an instance of this characteristic dual adjustment between definition and usage, whereby the usage informs the definition, which in turn guides extension of the usage"

[39] See FFF 64 quoted above; Goodman explicitly reinforced this point in a late interview (1995: p. 347).

[40] In an earlier version, Goodman explicitly wrote: "The sign '$=_{df}$' […] indicates a rule of translation between the formal system and ordinary discourse" (Goodman 1990: p. 77).

(FFF 66). Goodman seems to imply that commitments correspond to the explicandum ("usage"), that principles correspond to the definiens for the explicatum ("definition"), and that the process of mutual adjustments includes adjusting the extension of the explicandum as well as giving a constructional definition. But how exactly should we relate the method of RE to the structure of constructional definitions depicted in Fig. 2? What about the aspects that seem not present in FFF, most conspicuously the clarification of the explicandum and the distinction between explicandum$_1$ and explicandum$_2$?[41]

I propose to answer these questions with the help of the assumption, held implicitly by Goodman, that the set of commitments and the system of principles each characterize a concept (given the simplification that we deal with one concept only). The extension of our pre-theoretical concept of logical validity, for example, can then be seen as characterized by commitments to the validity of inferences, for example, the commitments that (1) is valid, but (3*) and (5*) are not; and the concept of logical validity given in a logical theory can be seen as determined by the principles of valid inference specified in this theory. Similarly, the extension of the everyday concept *fish* is characterized by commitments expressed in judgements such as "This is a fish" (pointing to an animal in a fish bowl), "Fish live in water" or "I don't eat fish" (when offered a shark steak). *Piscis* on the other hand is determined by a definition in biological terms, which reads, let us assume, something like "pisces are cold-blooded aquatic vertebrate having gills throughout life". Given such a relation of concepts to commitments and principles, the transition from SA to FFF can be interpreted as involving a shift in perspective. Whereas in SA, concepts and their extensions are in the foreground, FFF addresses these concepts indirectly through commitments and principles that characterize them.

We can now explain the connection between RE and constructional definition for the case of logical validity as follows. The explicatum is the concept of logical validity as defined in a logical theory and used, for example, in expressions of the form "$\Gamma \vdash \phi$" (which express that the inference from the set of premises Γ to the conclusion ϕ is valid); the definiens for the explicatum are the principles of validity given in this theory, say, rules of natural deduction or a model-theoretic definition of validity; the explicandum-term is "… is a valid inference"; the explicandum$_1$ is the pre-theoretical concept of logical validity as characterized by what we are committed to with respect to validity before we start the process of explicating; the explicandum$_2$, finally, is the (technical) concept of validity determined by the validity-commitments that are in agreement with the principles.

Therefore, what the theory of constructional definition describes as the result of adjusting the extension of "… is a valid inference", RE describes as the result of mutually adjusting commitments and principles of validity; and the criterion of extensional isomorphism is involved in assessing whether the commitments in effect agree with the principles. This description also shows that we need to distinguish between *initial* commitments and the commitments *resulting* from the process of developing a RE. The former relate to the pre-theoretical concept we start out with, the explicandum$_1$,

[41] These questions, as far as I know, have never been raised in the literature on RE, not even in Hahn's (2000) extensive discussion of SA.

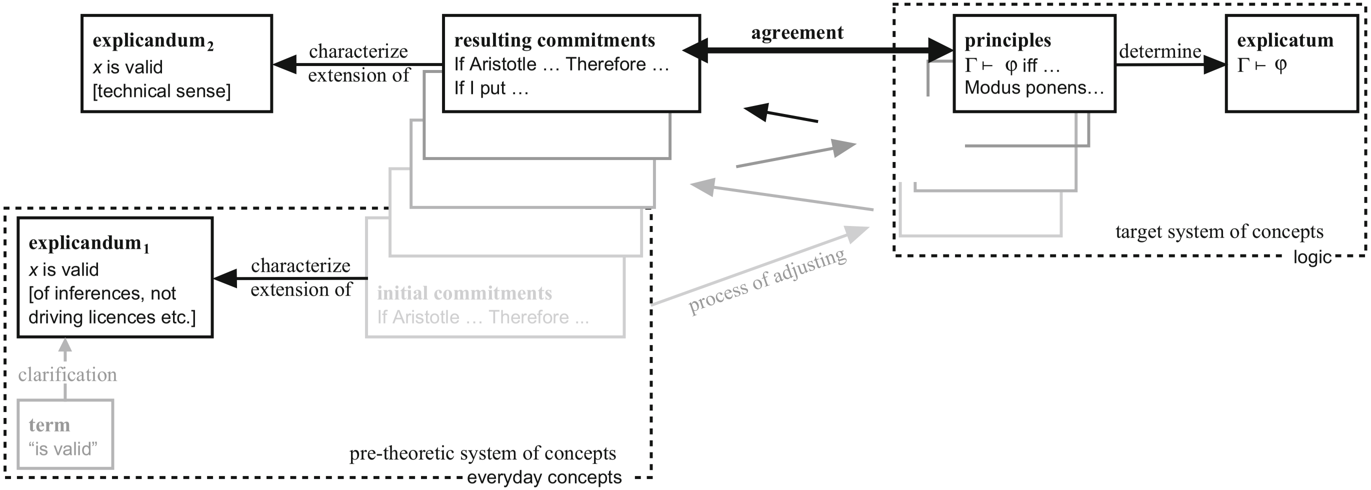

Fig. 3 Reflective equilibrium effecting a constructional definition of *valid*

the latter to the "trimmed and patched" explicandum$_2$. Figure 3 shows the structure of RE and its links to constructional definition and explication.

The picture drawn in Fig. 3 helps to clarify some crucial aspects of Goodman's approach to conceptual re-engineering. We can first note a shift of attention from Goodman's discussion of constructional definition in SA to his account in FFF. In Fig. 2, we distinguished two steps involved in giving a constructional definition: adjusting the explicandum and replacing it by the explicatum. Whereas in SA Goodman focused on the second step and the criterion of extensional isomorphism, he says virtually nothing about that in FFF and rather concentrates on the first step, the adjusting of the explicandum. As a consequence, Goodman no longer describes conceptual re-engineering as a linear process but as a process of iterative and bidirectional adjustments.[42] The result of one adjustment is the basis for the next, and adjustments can either adapt the commitments in light of the principles or the other way around.

This non-linear process shows, firstly, how the "trimming and patching" referred to in SA is implemented: we start out with an explicandum$_1$, then we come up with a tentative explicatum, investigate whether they are extensionally isomorphic, and if this is not the case we adjust either the explicatum or the explicandum until we have come up with an explicatum and an explicandum$_2$ which are extensionally isomorphic. This confirms the diagnosis (made at the end of Sect. 4) that the explicandum$_2$ is not introduced in advance of giving a constructional definition, but rather as a result of giving one. Importantly, the adjustments of the explicandum are embedded in a process of theory development. They must be guided by considerations of systematicity and must be steps towards a system of concepts which prove theoretically useful and similar enough to the original concepts.

Secondly, replacing the linear picture by a process of mutual adjustments paves the way for Goodman's claim that a RE constitutes a justification of *both* the principles and the commitments in agreement with them (FFF 64 quoted in Sect. 5.1). Such a claim would be implausible in the context of a linear method. That an explication or a constructional definition meets the conditions of adequacy can be interpreted as a justification for adopting the explicatum in place of the explicandum. This is a justification not for using, but rather for not using the explicandum. In the case of RE, this is different because RE is not meant as a justification for initial but rather for resulting commitments, and hence not for the pre-theoretical explicandum$_1$, but for the explicandum$_2$, which results from the process of conceptual re-engineering. The state of RE justifies adopting the adjusted explicandum$_2$ for theoretical purposes and replacing the explicandum by the explicatum in the resulting theory. Taking Goodman's example: when doing Geometry, we do not rely on the everyday concept *point* but on an adjusted concept of point (the explicandum$_2$), to be understood (roughly) as a dimensionless geometric element localized in space, and within the system of concepts of Geometry, this notion of point is replaced by an explicatum that is extensionally isomorphic to the explicandum$_2$ and defined as, for example, pair of intersecting lines.

[42] Nagel (2007) argues that this contrast constitutes a fundamental difference between explication and RE. If the analysis given here is correct, this overstates the case, because RE is less an alternative than a further development of the method of explication.

Furthermore, we are now in a position to clarify what is wrong with Goodman's characterization of commitments and principles as particular and general respectively. The structure depicted in Fig. 3 shows that the crucial difference is not related to the content of commitments and principles, but to the fact that the principles characterize concepts of the target system, explicata, whereas the commitments relate to explicanda, which are not part of the target system. Paradigmatic examples of theories of logical validity such as *Principia Mathematica* confirm this diagnosis. We can refer to, say, modus ponens as an element of a logical theory or as a commitment to the validity of all inferences of a certain form. In fact, such a commitment must be included in the resulting commitments if an RE with principles of classical logic is to be reached. The garden-variety understanding of RE as an interplay between particular and general should therefore be replaced. Conceptual re-engineering is rather about the relation of a practice to a theory, of established concepts to a theoretical system of concepts, and of commitments characterizing pre-theoretical concepts to principles determining concepts which are part of a theoretical system.

Finally, it is crucial to appreciate that the transition from explication and constructional definition to RE involves a switch from concepts to commitments and principles characterizing them, and that this switch is not merely a shift in emphasis. Although theoretical goals drive explications, the method of explication still focuses on definitions (or other characterizations) of individual concepts. RE, on the other hand, deals with a system of principles that constitutes a theory, thereby firmly placing conceptual re-engineering into the process of theory development (see also Hempel 2000). As we have seen, explicators very often face the challenge that they cannot simply draw on a pre-developed target theory, but need to contribute to theory development as well. In RE, this intertwining of conceptual re-engineering and theory development takes centre stage.

The points made in the preceding paragraphs have some interesting consequences for the interpretation of Carnap's methodological ideas. As noted in Sect. 3.4, Carnap's actual practice of explication, in contrast to his explicit methodology, aims at developing systems of concepts and occasionally proceeds in a non-linear way. That these features take central stage in Goodman's methodological proposals confirms the view that Goodman can be read as developing Carnap's ideas further. This, in turn, gives new plausibility to the view, recently defended by Carus, that Carnap was in fact moving towards a non-linear, "dialectical", method of philosophy. And it suggests that Goodman's ideas can be used to actually develop, as it is my strategy in this paper, a more detailed picture of how such a dialectical method could actually look like.[43]

5.4 Two additional aspects of reflective equilibrium

The preceding sections are not intended to describe a development that simply culminates in Goodman's account of RE in FFF. Although the method of RE as described so far effectively incorporates and extends the structure of explications and constructive definitions, my exposition of RE has already extrapolated a good deal from Good-

[43] Thanks to an anonymous reviewer for prompting me to say more about Carus's proposals.

man's explicit description in FFF and there is still more to be learned from the theories of constructional definition and explication. Specifically, the target state of RE has so far been characterized by just one criterion, namely the agreement of commitments and a system of principles. But our discussion of criteria of adequacy in the preceding sections calls for supplementing Goodman's account of RE by explicitly adding two further aspects.[44]

On the one hand, considerations of similarity have important implications for RE. We have seen that Goodman's theory of constructional definition actually splits Carnap's similarity requirement into two conditions, one covered by extensional isomorphism and one which applies to the relation between the explicandum$_1$ and the explicandum$_2$. The latter places constraints on the "trimming and patching" and consequently on the admissible adjustments in the process of developing a RE. As explained in Sect. 4, the constraint is meant to ensure that explications do not simply change the subject. In its most general form, it can be framed as Carnap's requirement that the explicatum can be used in place of the explicandum in relevant contexts. If we apply this idea to RE, it amounts to the constraint that the process of mutual adjustments does not implement revisions so drastic that we end up with principles which, for example, do not count as a theory of valid inference any more. Principles declaring all inferences drawn by certain people to be "valid" might perhaps provide an account of authorship or authority but certainly not of logical validity. Justification by RE therefore requires that the resulting commitments (and indirectly the resulting principles) adequately respect initial commitments.[45]

On the other hand, virtues of theories play just as important a role in RE as they do in explication. Although Goodman only incidentally refers to desired virtues of theories in the passages of FFF that deal with RE,[46] the points made in the context of Carnap's method of explication clearly remain valid. We have to recognize that there are a variety of desiderata, some of which are generally relevant theoretical virtues such as fruitfulness, simplicity, exactness and wide range of application,[47] while others are specific, for example, to theories of logical validity, such as amenability to rigorous proof and decidability. Further desiderata are constituted by the functions the explicatum is expected to perform in the target theory and the problems it is expected to solve. Moreover, the theoretical virtues are partly antagonistic (e.g. precision may

[44] In the literature on RE, these points are typically neglected, except in the work of Elgin (esp. 1996); see also Baumberger and Brun (2016).

[45] As I introduce it here, the requirement of respecting antecedent commitments is actually Carnap's criterion of similarity transposed into the framework of reflective equilibrium. The resulting conception of reflective equilibrium is therefore not in danger of being inherently more "conservative" than explication as suggested by Dutilh Novaes and Geerdink (2017).

[46] He mentions "convenience", "theoretical utility" (FFF 66), "maximum coherence and articulation", "economy", "resultant integration" (FFF 47, referred to in FFF 65n2), and via the link to SA (in FFF 67n3) we can add "simplicity" of conceptual resources (SA ch. III.3). Goodman worked extensively on two theoretical virtues, confirmation and conceptual simplicity (in FFF and SA respectively).

[47] Since RE is often understood as a coherentist account of justification, one may wonder whether coherence should not replace the list of theoretical virtues or at least be included in it. However, the first option is too restrictive (excluding e.g. numerical precision); the second is unsuitable because some crucial aspects of coherence (consistency of commitments and principles as well as agreement of commitments and principles) are built into RE as necessary conditions.

conflict with scope of application) and trade-offs must be worked out in the light of the cognitive aims that guide the construction of a system of principles.[48] We therefore cannot expect that the resulting system exhibits all relevant theoretical virtues to a maximal degree. To express this, I say that systems of principles are expected to "do justice to" (rather than, e.g. "realize") desiderata for theories.

For RE, doing justice to desiderata for theories is crucial because it ensures that the principles provide a *systematic* account, not merely a list of commitments. And it plays a key role in the process of adjusting: if principles contribute to the realization of virtues such as simplicity and wide scope, this constitutes a (defeasible) reason to adjust commitments. Proponents of classical logic may, for example, be ready to dismiss the commitment that (5*) is invalid on the grounds that the material conditional (for which it is characteristic that (2) and (6), but not (4*), are rules of valid inference) provides a simple, elegant and uniform account of indicative conditional sentences.

The reference to theoretical virtues reinforces the insight that re-engineering concepts for theoretical purposes must always be seen in the wider context of theory development. Virtues *of theories* are indeed the driving force behind the methods of explication, constructional definition and RE, because they motivate the transition to a theoretically more suitable system of concepts in the first place.

6 Prospects for methods of conceptual re-engineering

The last two points actually go to the core of the methods of conceptual re-engineering. Explication, constructional definition and RE are driven by two forces: respecting initial commitments or similarity to the original explicandum on the one side, and doing justice to the desiderata which drive the theoretical development in the context at hand on the other. Between these two, an adequate equilibrium must be sought—hence we have a second reading of the metaphor "reflective equilibrium", which is usually interpreted as referring to the agreement between commitments and principles. This core provides the systematic reason for interpreting Carnap and Goodman as contributing to a common philosophical programme of conceptual re-engineering. It is also the motivation for subsuming their methods under the label "conceptual re-engineering for theoretical purposes". "Engineering" highlights that these are methods of explicit and intentional concept development, in contrast to, for example, spontaneous language change. "Re-engineering" alludes to "reconstruction" and to Carnap's "language engineering" (1963a: p. 66), and expresses that the conceptual engineering is guided by pre-existing concepts. The qualification "for theoretical purposes", finally, emphasizes that Carnap and Goodman develop methods for theory development in science or philosophy, in contrast to conceptual re-engineering driven, for example, by purely practical considerations in legal contexts, concerns of political correctness or a political agenda which hopes to benefit from newspeak.

My discussion uncovered interconnections between explication, constructional definition and RE. On the one hand, Goodman was presented as addressing two limitations

[48] Just as in the case of explication and constructional definition, the necessity of trade-offs is also a reason for expecting and actually welcoming that RE allows for justifying alternative systems.

of Carnap's method of explication: the theory of constructional definition overcomes the focus on individual concepts by dealing with systems of concepts, and RE replaces the linear structure of explications by an iterative and bidirectional method. On the other hand, my discussion of Carnap's criteria of adequacy for explications highlighted aspects of conceptual re-engineering which Goodman's explanation of RE does not address. Specifically, it made clear that conceptual re-engineering is driven by and accountable to considerations of theoretical virtues and similarity to the original concept. This called for supplementing the two core ideas of RE—adjustment and agreement—by the two requirements of doing justice to desiderata for theories and respecting initial commitments.

On the whole, the analysis in this paper shows that combining Carnap's and Goodman's contributions to methodology has considerable potential, and, specifically, suggests how their ideas can be incorporated into an account of RE that deals with conceptual re-engineering as a component of theory development. However, we are still quite far from a full-fledged account of RE based on the ideas of Carnap and Goodman. All the main aspects of RE call for more thorough investigation. We need, for example, a deeper analysis of the relation of agreement between commitments and principles, a detailed specification of the process of adjustments and an explanation of its epistemic significance, an explanation of how the relative weight of commitments and principles is to be understood, and a more informative explanation of what doing justice to desiderata and respecting initial commitments amounts to. Moreover, if we take seriously that RE deals with principles which constitute theories, further questions arise relating to, for example, the credibility and accuracy of theories. Finally, I bracketed the issue of background theories, the support of which constitutes another aspect of the justification of commitments and principles in (so called "wide") RE. Addressing these issues lies well beyond the scope of this paper.[49] It calls for reconnecting the discussion about explication and RE to the debates about theory choice and epistemic justification.

Acknowledgements The research for this paper is part of the project "Reflective Equilibrium – Reconception and Application" (Swiss National Science Foundation Project 150251). It draws on material I wrote while I was a visiting scholar at Harvard University (supported by SNSF Grant 125823). Earlier versions have been presented in Berne, Frankfurt a.M., Prague and Zürich. I would like to thank the audiences, two anonymous reviewers of this journal and especially Christoph Baumberger, Claus Beisbart, Catherine Elgin, Leon Geerdink, Jaroslav Peregrin, Geo Siegwart, Vladimír Svoboda and Nicolas Wüthrich for helpful comments and discussions.

References

Bar-Hillel, Y., & Carnap, R. (1953). Semantic information. *British Journal for the Philosophy of Science*, *4*, 147–57.

Baumberger, C. (forthcoming). Explicating objectual understanding. Taking degrees seriously. *Journal for General Philosophy of Science*.

[49] Elgin has undertaken a project of developing a comprehensive account of RE starting from Goodman's ideas (Elgin 1983:ch. X; 1996), and elsewhere I have sketched the structure of an account of RE that incorporates the points made in the preceding sections (Brun 2014a, b; Baumberger and Brun 2016).

Baumberger, C., & Brun, G. (2016). Dimensions of objectual understanding. To appear in S.R. Grimm, C. Baumberger, S. Ammon (Eds). *Explaining understanding. New perspectives from epistemology and philosophy of science* (pp. 165–189). New York: Routledge.

Brun, G. (2014a). Reconstructing arguments: Formalization and reflective equilibrium. *Logical Analysis and History of Philosophy*, *17*, 94–129.

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

Brun, G. (2016). Explication as a method of conceptual re-engineering. *Erkenntnis*, *81*, 1211–1241.

Cappelen, H. (forthcoming). *Fixing Language. An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Carnap, R. (1945). The two concepts of probability. *Philosophy and Phenomenological Research*, *5*, 513–532.

Carnap, R. (1956) [1947]. *Meaning and necessity. A study in semantics and modal logic* (2nd ed.). Chicago: University of Chicago Press.

Carnap, R. (1962) [1950]. *Logical foundations of probability* (2nd ed.). Chicago/London: University of Chicago Press/Routledge and Kegan Paul. Referenced as LFP.

Carnap, R. (1963a). Intellectual autobiography. In P. A. Schilpp (Ed.), *The philosophy of Rudolf Carnap* (pp. 3–84). La Salle: Open Court.

Carnap, R. (1963b). Replies and systematic expositions. In Schilpp, P. A. (Ed.). *The philosophy of Rudolf Carnap* (pp. 859–1013). La Salle: Open Court. Referenced as RSE.

Carnap, R. (2003) [1928/34]. The logical structure of the world. In: Pseudoproblems in philosophy. Chicago/La Salle: Open Court.

Carus, A. W. (2007). *Carnap and twentieth-century thought. Explication as enlightenment*. Cambridge: Cambridge University Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Cohnitz, D., & Rossberg, M. (2006). *Nelson Goodman*. Montreal/Kingston: McGill-Queen's University Press.

Daniels, N. (1980). On some methods of ethics and linguistics. *Philosophical Studies*, *37*, 21–36.

Daniels, N. (2016). Reflective equilibrium. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/win2016/entries/reflective-equilibrium/.

DePaul, M. R. (2011). Methodological issues reflective equilibrium. In C. Miller (Ed.), *The continuum companion to ethics* (p. lxxv-cv). London: Continuum.

Douglas, H. (2013). The value of cognitive values. *Philosophy of Science*, *80*, 796–806.

Dutilh Novaes, C., & Reck, E. (2017). Carnapian explication, formalisms as cognitive tools, and the paradox of adequate formalization. *Synthese*, *194*, 195–215.

Dutilh Novaes, C., & Geerdink, L. (2017). The dissonant origins of analytic philosophy. Common sense in philosophical methodology. In S. Lapointe & C. Pincock (Eds.), *Innovations in the history of analytical philosophy* (pp. 69–102). London: Palgrave Macmillan.

Elgin, C. Z. (1983). *With reference to reference*. Indianapolis: Hackett.

Elgin, C. Z. (1996). *Considered judgment*. Princeton: Princeton University Press.

Elgin, C. Z. (1997). Volume introduction. In C. Z. Elgin (Ed.). *Nominalism, constructivism, and relativism in the work of Nelson Goodman* (pp. xiii–xvii). New York: Garland. (= The Philosophy of Nelson Goodman. Selected Essays; Vol. 1).

Goodman, N. (1963) [1956]. The significance of Der logische Aufbau der Welt. In P. A. Schilpp (Ed.), *The philosophy of Rudolf Carnap* (pp. 545–58). La Salle: Open Court.

Goodman, N. (1968). *Languages of art. An approach to a theory of symbols*. Indianapolis: Bobbs-Merrill.

Goodman, N. (1972a). *Problems and projects*. Indianapolis/New York: Bobbs-Merrill.

Goodman, N. (1972b) [1970]. Seven strictures on similarity. In Goodman 1972a: pp. 437–446.

Goodman, N. (1977) [1951]. *The structure of appearance* (3rd ed.). Dordrecht/Boston: Reidel. (1st ed. 1951. Cambridge, MA: Harvard University Press. 2nd ed., 1966, Bobbs-Merrill, Indianapolis.) Referenced as SA.

Goodman, N. (1983) [1954]. *Fact, fiction, and forecast* (4th ed.). Cambridge, MA: Harvard University Press. Referenced as FFF.

Goodman, N. (1990) [1941]. *A study of qualities*. New York: Garland.

Goodman, N. (1995). Gewißheit ist etwas ganz und gar Absurdes. Karlheinz Lüdeking sprach mit Nelson Goodman. *Kunstforum*, *131*, 342–347.

Griffin, J. (2008). *On human rights*. New York: Oxford University Press.

Haas, G. (2015). *Minimal verificationism. On the limits of knowledge*. Berlin: de Gruyter.

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

Hanna, J. F. (1968). An explication of explication. *Philosophy of Science*, *35*, 28–44.

Harper, W. L. (1981). A sketch of some recent developments in the theory of conditionals. In W. L. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs. Conditionals, belief, decision, chance, and time* (pp. 3–38). Dordrecht: Kluwer.

Hegselmann, R. (1985). *Formale Dialektik. Ein Beitrag zu einer Theorie des rationalen Argumentierens*. Hamburg: Meiner.

Hellman, G. (1977). Introduction. In SA XV–XLVII.

Hellman, G. (1978). Accuracy and actuality. *Erkenntnis*, *12*, 209–228.

Hempel, C. G. (1952). *Fundamentals of concept formation in empirical science*. Chicago: University of Chicago Press.

Hempel, C. G. (1953). Reflections on Nelson Goodman's the structure of appearance. *The Philosophical Review*, *62*, 108–116.

Hempel, C. G. (1970). Aspects of scientific explanation. *Aspects of scientific explanation and other essays in the philosophy of science* (pp. 331–496). New York: Free Press.

Hempel, C. G. (2000) [1988]. On the cognitive status and the rationale of scientific methodology. In *Selected philosophical essays* (pp. 199–228). Cambridge: Cambridge University Press.

Justus, J. (2012). Carnap on concept determination. Methodology for philosophy of science. *European Journal for the Philosophy of Science*, *2*, 161–179.

Kant, I. (1998). [1781/1787]. Critique of pure reason. Cambridge: Cambridge University Press.

Kantorovich, A. (1993). *Scientific discovery*. Logic and tinkering. New York: State University of New York Press.

Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. In *The essential tension*. Selected Studies in Scientific Tradition and Change. Chicago: University of Chicago Press, pp. 320–39.

Kuipers, T. A. F. (2007). Introduction. Explication in philosophy of science. In Kuipers, T. A. F. (Ed.). *Handbook of the philosophy of science. Focal issues* (pp. vii–xxiii). Amsterdam: Elsevier.

Leitgeb, H. (2017). *The stability of belief. How rational belief coheres with probability*. Oxford: Oxford University Press.

Lutz, S. (2012). Artificial language philosophy of science. *European Journal for the Philosophy of Science*, *2*, 181–203.

Maher, P. (2007). Explication defended. *Studia Logica*, *86*, 331–341.

Maher, P. (2010). *What is probability?* MS. http://patrick.maher1.net/preprints/pop.pdf. Accessed April 17, 2016.

Mäkinen, J., & Kakkuri-Knuuttila, M.-L. (2013). The defence of utilitarianism in early rawls. A study of methodological development. *Utilitas*, *25*, 1–31.

Martin, M. (1973). The explication of a theory. *Philosophia*, *3*, 179–199.

Mikhail, J. (2010). Rawls' concept of reflective equilibrium and its original function in a theory of justice. *Washington University Jurisprudence Review*, *3*, 1–30.

Miller, R. B. (2000). Without intuitions. *Metaphilosophy*, *31*, 231–250.

Nagel, J. (2007). Epistemic intuitions. *Philosophy Compass*, *2*, 792–819.

Olsson, E. J. (2015). Gettier and the method of explication. A 60 year old solution to a 50 year old problem. *Philosophical Studies*, *172*, 57–72.

Pap, A. (1949). The philosophical analysis of natural language. *Methodos*, *1*, 344–369.

Pinder, M. (2017). Does experimental philosophy have a role to play in carnapian explication? *Ratio*, https://doi.org/10.1111/rati.12164.

Prior, A. N. (1960). A runabout inference-ticket. *Analysis*, *21*, 38–39.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA.: MIT Press.

Quine, W. V. O. (1980) [1951]. Two dogmas of empiricism. In *From a logical point of view. Nine Logico-philosophical essays*. (2nd Ed., pp. 20–46). Cambridge, MA: Harvard University Press.

Rawls, J. (1950). *A study in the grounds of ethical knowledge. Considered with reference to judgments on the moral worth of character*. PhD dissertation, Princeton University.

Rawls, J. (1999a) [1974–5]. The independence of moral theory. In *Collected papers* (pp. 286–302). Cambridge, MA: Harvard University Press.

Rawls, J. (1999b) [1951]. Outline of a decision procedure for ethics. In *Collected papers* (pp. 1–19). Cambridge, MA: Harvard University Press.

Rawls, J. (1999c). *A theory of justice* (revised ed.). Cambridge, MA: Belknap Press.

Rawls, J. (1999d) [1955]. Two concepts of rules. In *Collected papers* (pp. 20–46). Cambridge, MA: Harvard University Press.

Reck, E. (2012). Carnapian explication. A case study and critique. In P. Wagner (Ed.), *Carnap's ideal of explication and naturalism* (pp. 96–116). Basingstoke: Palgrave Macmillan.

Russell, B. (1954). *Mysticism and logic and other essays*. Harmondsworth: Penguin.

Russell, B. (1993) [1914]. *Our knowledge of the external world as a field for scientific method in philosophy*. London: Routledge.

Scanlon, T. M. (2008). *Moral dimensions. Permissibility, meaning, blame*. Cambridge, MA: Harvard University Press.

Scheffler, I. (1954). On justification and commitment. *Journal of Philosophy*, *51*, 180–190.

Shepherd, J., & Justus, J. (2015). X-Phi and Carnapian explication. *Erkenntnis*, *80*, 381–402.

Schupbach, J. N. (2017). Experimental explication. *Philosophy and Phenomenological Research*, *94*, 672–710.

Stein, E. (1996). *Without good reason*. The Rationality debate in philosophy and cognitive science. Oxford: Clarendon Press.

Strawson, P. F. (1963). Carnap's views on constructed systems versus natural languages in analytic philosophy. In P. A. Schilpp (Ed.), *The philosophy of Rudolf Carnap* (pp. 503–18). La Salle: Open Court.

Tarski, A. (1983) [1933]. The concept of truth in formalized languages. *Logic, semantics, metamathematics* (2nd Ed., pp. 152–278). Hackett: Indianapolis.

Tarski, A. (2002) [1936]. On the concept of following logically. *History and Philosophy of Logic*, *23*, 155–196.

Tersman, F. (1993). *Reflective equilibrium*. An essay in moral epistemology. Stockholm: Almqvist and Wiksell.

Vermeulen, I. (2013). *Words matter. a pragmatist view on studying words in first-order philosophy*. PhD thesis, University of Sheffield. http://etheses.whiterose.ac.uk/id/eprint/4759.

Whitehead, A. N. (1919). *An enquiry concerning the principles of natural knowledge*. Cambridge: Cambridge University Press.