Synthese
DOI 10.1007/s11229-015-0793-2

S.I.: CARNAP ON LOGIC

Value concepts (1958)

Rudolf Carnap

Received: 22 January 2015 / Accepted: 1 June 2015

Abstract Carnap wrote a continuation of his reply to Kaplan (§32 of Carnap's replies in the 1963 Schilpp volume), which would, however, have made that reply, already by far the longest in the book, too long. So he set aside his projected notes for a continuation to serve as the basis for a separate paper, which he never got around to writing. It is transcribed here from his shorthand and translated into English, with some introductory notes to provide a little context.

Keywords Carnap · Decision theory · Instrumental vs. substantive rationality · Rationality · Post-Kantian value theory

Introductory remarks (A.W. Carus)

As Rudolf Carnap was revisiting the final section (on values) of his replies to critics in the Schilpp volume on *The Philosophy of Rudolf Carnap* (Carnap 1963), he realized that the suggestions about formalizing value concepts toward the end of that reply were rather vague, and he decided to spell out something a little more definite along those lines. He wrote several pages of shorthand over a couple of days, under the heading *Wertbegriffe* (value concepts), then decided that the new material would make this section (already by far the bulkiest of all the replies) disproportionately long. Still, he liked the approach he had sketched, and decided to keep the notes for a separate paper, which never materialized.

Editor's note: The beginning of a new page in Carnap's manuscript is indicated here by the new page number in square brackets.

The resulting shorthand fragment remained in his papers at the Archive for Scientific Philosophy in Pittsburgh[1] and has, as far as I can tell, not been discussed. It follows below, transcribed and translated into English from the odd mishmash of Anglified German in which Carnap took down shorthand notes in his later years.[2] These introductory notes will sketch some relevant context and briefly indicate why this fragment is of more than antiquarian interest.

The final section (§32) of Carnap's replies of which this fragment was intended to be a part was Carnap's reply to Abraham Kaplan, a former graduate student at the University of Chicago.[3] This reply was Carnap's only extended foray into the logic of normative and value statements.[4] It was largely ignored by philosophers of meta-ethics, perhaps because they discerned that §32 closely resembled the exposition of the logic of normative statements given by Richard Hare in *The Language of Morals* a few years earlier.[5] Both Carnap and Hare were non-cognitivists who wanted to account for the obvious and very extensive factual or descriptive components in normative sentences without conflating the two categories; both built on G. E. Moore's "naturalistic fallacy" argument and on Stevenson's *Ethics and Language*. But Hare had put forward his account in the style of Oxford ordinary-language philosophy. Even if Carnap had looked at *The Language of Morals* (which is actually cited in the Kaplan paper he was replying to)[6] it seems unlikely he would have appreciated the close similarity to his own account of normative language, descriptive language, and their inferential interrelations. To moral philosophers, on the other hand, the resemblance would have been more obvious, and they might well have thought it superfluous to respond to Carnap when Hare was already at the center of attention.

Hare was not a doctrinaire or off-the-shelf ordinary-language philosopher. He appealed more to the *functional* difference between descriptive and normative statements (he called the latter "action-guiding", with moral statements a tiny subclass—just the most general ones) than to ordinary usage itself. Still, he opened himself up to the criticism (e.g. by his Oxford successor Bernard Williams (1985)) that the heterogeneity of actual spoken language calls the simple partition of all sentences into

[1] It is located in the Carnap papers (RC) at 89-14-01. A scan of the original shorthand manuscript is also available online at http://digital.library.pitt.edu/u/ulsmanuscripts/pdf/31735061815522.pdf.

[2] A transcription of the original "German" text is available (though this would undoubtedly have embarassed Carnap somewhat) at http://awcarus.com/2015/04/carnap-on-value-concepts/.

[3] And later a colleague of Carnap's at UCLA; Kaplan's (1991) vivid memoir of Carnap as a teacher and mentor at the University of Chicago is full of affectionate admiration.

[4] Some earlier writings (§152 of Carnap 1928; Carnap 1934) on the subject were much briefer and less systematic, but have nonetheless inspired more commentary than Carnap (1963); see e.g. Mormann (2006, 2010), Uebel (2010) and Richardson (2007). Still, §32 has not gone unnoticed (e.g. Uebel 2005, esp. p. 769, and Dreben 1995).

[5] Hare's book was published in 1952, and immediately attracted widespread attention; Kaplan's critique of Carnap on values, citing Hare (see footnote 6 below), was probably written during 1955, and Carnap's reply the year after that. The Schilpp volume on Carnap remained unpublished until 1963, however, as a new publisher for the series had to be found.

[6] Kaplan mentions Hare (1952) as the latest in a series of attempts by the "British school" to distinguish the cognitive from the normative components in sentences, an effort he thinks both mysterious and completely at odds with logical empiricism. It would perhaps repay historical excavation to explore why he might have held this opinion.

descriptive and normative into question. There is no basis in ordinary language *itself* for imposing such a schema; it has to be imported from outside.

This is where it would have helped if Carnap's exposition had attracted a little more attention, as an alternative or complement to Hare's. For Carnap, ordinary usage lacked the authoritative status it had for Strawson, Williams, Hare, or even, in a different way, for Quine. In Carnap's own scattered remarks on this theme, he often echoed Fregean sentiments about the misleading nature of ordinary language. Unlike Frege, of course, he did not think there was an "underlying" structure of thoughts residing in a third realm; "Carnap rejects Frege's assumption of a common store of logically interrelated thoughts expressed by the sentences of colloquial language and perspicuously expressible by sentences couched in the framework of Begriffsschrift". (Ricketts 2004, p. 191) Carnap's version thus has two possible advantages over Hare's: first, it is more consistent with Hare's own (early) aim of developing a logic for normative (and thus moral) language, as it does not conflate that task with the completely different one of extracting from ordinary language the distinctions embedded in it (cf. Uebel 2005, esp. p. 769). So it is not vulnerable to the critique that it fails to map onto ordinary language, while it can still legitimately claim to *explicate* (Carnap 1950, pp. 1–6) certain distinctions that appear to play a central role in aspects of ordinary life. Secondly, Hare greatly complicated the reception of his framework for normative language by proceeding, before long, to build an ambitious utilitarianism on its foundation. This later development, for good or ill, distracted many from the more basic question of the underlying account of normative language in *The Language of Morals*. Carnap, as the fragment below makes evident, was not ultimately a utilitarian or even, perhaps, a consequentialist.

This will surprise many readers, as Carnap has often been seen, insofar as any general framework of values and rationality has been attributed to him at all, as a—perhaps somewhat heterodox—proponent of Bayesian decision-theoretic rationality (e.g. Earman 1993; Gower 1997). And it is true that, within the realm of inductive logic and its wide range of practical applications, this was very much his view. What the present document makes evident, however, is that he saw inductive value functions, defined by axioms of induction and the choice of an inductive method, as *partial* value functions, i.e. as guiding choices only over a restricted range of an individual's (or a society's) overall priorities.

Opinions will differ about how to characterize the view Carnap sketches. If a minimal Kantianism is suggested by the distinction between "purely valuational" criteria of rationality for moral value functions (p. [6][7]) and instrumental criteria for partial value functions (which may be regarded as an explication of Kant's distinction between *Vernunft* and *Verstand*), it is evidently a more rarefied, and less Rousseau-oriented, Kantianism from those worked out in more laborious detail by, e.g. Rawls or Habermas.[8] Still, it is worth noting that Carnap himself rejects a certain kind of consequentialism in this document:

[7] Page references to Carnap's manuscript, in square brackets, are to the original document; in the translation below, they are embedded in the text in square brackets.

[8] To which it was compared, though in ignorance of the present document, by Carus (2007, pp. 297–309); see also Carus (this volume). A fascinating and surprising parallel between Rawls and Carnap is drawn in the concluding paragraphs of Dreben (1995).

> Assume X is perfectly rational at time t and chooses action a in A_X. Then it is nonetheless still possible for a *not to be an optimum* with respect to V_X [X's comprehensive value function]. It could be that an action a' is better than a with respect to V_X, due to certain circumstances not known to X at the time of the action. It could even be that the objectively better, i.e. more successful action a' would not be rational for X. As emphasized elsewhere (§[26.IV][9]), rationality is not to be determined by success. (p. [10])

Carnap refers here to the passages from his 1963 replies regarding the use of experience in the choice of axioms for inductive logic, and of inductive methods, so as to ensure that the choices they lead to are rational.[10] Here the analogy between the partial value functions bearing on the choice of inductive axioms and methods, on the one hand, and comprehensive or moral value functions on the other, becomes explicit, with respect to the relevance of experience to the respective choices. The analogy has limits; while instrumental rationality may *constrain* substantive (moral) rationality, in this view, it does not determine it; the "purely valuational" criteria Carnap invokes (p. [6] of the document below) ultimately govern the choice of values, and in this respect Carnap remains faithful to Kant.[11]

The overall view sketched by Carnap has some potentially attractive features. It combines a Bayesian decision-theoretic rationality at the cognitive (or more broadly instrumental) level with a kind of minimally Kantian substantive rationality at the level of ultimate values, without claiming (like Kant and some later Kantians) to be able to determine a single, unique highest principle of morality. There is a striking parallel between this idea and the "relativized a priori", as Michael Friedman has called it, of which different versions are suggested in Poincaré, Schlick, early Reichenbach, Cassirer, and Carnap. Just as (Kantian) unique synthetic a priori knowledge is relativized by these figures to different historical epochs or human purposes, so the (Kantian) unique categorical imperative is relativized by Carnap, in the fragment published here, to the many different fundamental values that prevail in different contexts and cultures. Not only does this conception leave room for value pluralism, then, but it clearly subordinates instrumental rationality to ultimate values in a way that has

[9] All references within Carnap's manuscript are to sections of his replies or others' papers in the Schilpp volume (Carnap 1963), for which the manuscript was originally intended.

[10] That he is referring to this passage is reinforced by other references back to it in the published text, e.g. "I do not share the widespread view that the rationality of an inductive method depends upon factual knowledge, say, its success in the past. I think that the question of rationality must be answered by purely a priori considerations (see my comments. . . in §26(IV)". (Carnap 1963, p. 981) The passages referred to here are quoted in Carus (this volume).

[11] It has been suggested that the constraints thus placed on possible "highest principles of morality" are "merely formal", and have no substantive bite. But it seems that Carnap is in no worse a position here than traditional Kantians who embrace the categorical imperative or some modernized version of it. For it is widely admitted that the categorical imperative is itself too abstract and "formal" to be applied to any concrete situation; it is in need, when it comes down to real life, of supplementation by the normative equivalent of "coordination rules". How are Carnap's constraints on the selection of such "highest principles" from the infinite set of candidate principles—which require the selection of a particular substantive principle in that set, arising from specific human purposes and ideals—more "formal" than that?

eluded some well-known attempts to conjoin these different components or levels of rationality.[12]

Carnap's strongest argument against deriving "perfect" rationality (at least) from successful outcomes comes in his final paragraph (though the connection is not made explicit):

> "More rational," whether applied to different periods or to two possible behaviors of the same person in the same period, cannot very well be exactly defined. Roughly speaking, a behavior is more rational than another when it comes closer to perfectly rational behavior. But since deviations from perfectly rational behavior are possible in completely different ways, e.g. in the ways mentioned above... and within each of these once again in different ways, it is hardly possible to decide without an arbitrary convention under what conditions a deviation in one way should be considered equal to a deviation in another way. (p. [10])

This impossibility of comparing, let alone measuring, different deviations from "perfect rationality" is in fact an immediate consequence of the sharp distinction between the criteria for determining instrumental (or partial) rationality from those governing substantive (comprehensive) rationality. If values are chosen by standards that are merely constrained (and not determined) by instrumental considerations, then distance from overall ("perfect") rationality would be arbitrary even if (as Carnap did not believe) *instrumental* rationality were only a matter of learning from experience or of past success.

It is both surprising and admirable that Carnap was so bluntly honest with himself about the consequences of his conception of rationality. For of course he was notoriously an advocate of quantitative concepts; he thought that psychology, for instance, would have to become more quantitative to be more scientific. And we find him admitting, here, that a quantitative measure of moral value functions is not feasible. It is probably not an accident that this fragment ends where it does, or that it was not ultimately picked up again and worked out. For while Carnap was honest enough to put down the words just quoted, the conclusion expressed in them must have been unwelcome to him.

Unburdened by this prejudice, we can appreciate the fragment for what it *does* suggest: a principled way of integrating instrumental and substantive rationality into a single coherent framework.[13] It casts an interesting light on Carnap's long years of struggle with inductive logic to know that he saw it as having a place within such a comprehensive conception of human thought and action. It casts an especially interesting light on Carnap's various remarks about the practical applicability of inductive logic, "probability as a guide in life", and reveals that they were not merely passively echoing Condorcet, Laplace, and the positivist tradition (let alone the English tradition of

[12] In Habermas, for instance, the weak coordination of instrumental, hermeneutic, and communicative rationalities and the lack of clarity about which form of ultimate meta-rationality is to govern any such coordination; in Rawls, the problematic relation between the "reasonable" and the "rational", and again, of the meta-reason that adjudicates between their respective scopes.

[13] Which is worked out in a little more detail in Carus (this volume).

Butler, Moore, and Keynes), but were rooted in a deeper and more complex—perhaps minimally Kantian—conception that was under constant re-examination.

Acknowledgments I am grateful to the Archive of Scientific Philosophy, Hillman Library, University of Pittsburgh, for permission to publish this document; all rights are reserved. I am especially grateful to Brigitta Arden for her assistance with a few questionable transcriptions, drawing on her extensive experience with Carnap's shorthand manuscripts. Florian Steinberger, Rick Creath, and Michael Friedman provided useful comments and feedback at various points in the evolution of the introductory comments; Thomas Uebel, Christian Damböck, and two anonymous reviewers provided very helpful feedback on the final draft. I am grateful to Georg Schiemer and the editors of *Synthese*, Otávio Bueno and Gila Sher, for their willingness to publish this fragment and their encouragement in seeing the idea through to completion.

References for Introductory Remarks

Carnap, R. (1928). *Logischer aufbau der welt*. Leipzig: Meiner.

Carnap, R. (1934). Theoretische fragen und praktische entscheidungen. *Natur und Geist*, *2*, 257–260.

Carnap, R. (1950). *Logical foundations of probability*. Chicago: University of Chicago Press.

Carnap, R. (1963). Replies and systematic expositions. In P. Schilpp (Ed.), *The philosophy of Rudolf Carnap* (pp. 859–1013). LaSalle, IL: Open Court.

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

Carus, A.W. (this volume). Carnapian rationality. Synthese, this issue.

Dreben, B. (1995). Cohen's Carnap, or subjectivity is in the eye of the beholder. In K. Gavroglu, J. Stachel, & M. W. Wartofsky (Eds.), *Science, politics, and social practice* (pp. 27–42). Dordrecht: Kluwer.

Earman, J. (1993). Carnap, Kuhn, and the philosophy of scientific methodology. In P. Horwich (Ed.), *World changes: Thomas Kuhn and the nature of science* (pp. 9–36). Cambridge, MA: MIT Press.

Gower, B. (1997). *Scientific method: An historical and philosophical introduction*. London: Routledge.

Hare, R. M. (1952). *The language of morals*. Oxford: Oxford University Press.

Kaplan, A. (1991). Rudolf Carnap. In E. Shils (Ed.), *Remembering the University of Chicago: Teachers, scientists, and scholars* (pp. 32–41). Chicago: University of Chicago Press.

Mormann, T. (2006). Werte bei Carnap. *Zeitschrift für philosophische Forschung*, *60*, 169–189.

Mormann, T. (2010). Wertphilosophische abschweifungen eines logischen empiristen: Der Fall Carnap. In A. Siegesleitner (Ed.), *Logischer empirismus, werte, und moral: Eine neubewertung* (pp. 81–102). Vienna: Springer.

Richardson, A. (2007). Carnapian pragmatism. In M. Friedman & R. Creath (Eds.), *The Cambridge Companion to Carnap* (pp. 295–315). Cambridge: Cambridge University Press.

Ricketts, T. (2004). Frege, carnap, and quine: Continuities and discontinuities. In S. Awodey & C. Klein (Eds.), *Carnap brought home: The view from jena* (pp. 181–202). LaSalle, IL: Open Court.

Uebel, T. (2005). Political philosophy of science in logical empiricism: The left Vienna Circle. *Studies in History and Philosophy of Science*, *36*, 754–773.

Uebel, T. (2010). 'BLUBO-metaphysik': Die verwerfung der werttheorie des Südwestdeutschen neukantianismus durch Carnap und Neurath. In A. Siegesleitner (Ed.), *Logischer empirismus, werte, und moral: Eine neubewertung* (pp. 103–129). Vienna: Springer.

Williams, B. (1985). *Ethics and the limits of philosophy*. Cambridge, MA: Harvard University Press.

Value Concepts (a shorthand manuscript by Rudolf Carnap, transcribed and translated by A.W. Carus)

Value concepts and rational agent First written to supplement my reply to Kaplan in the Schilpp volume. But that would have got too long. So better *as a basis* ***for a later paper!***

21 February 58

Value Concepts

Relatively to a value system

Let *V* be a *value function* (It is not assumed that there is a person whose value function is V.) This means that for every possible history of the world W, V(W) is a real number. Since only the differences among values of V matter, in the following definitions, two value functions V und V' that differ only by a constant (for every W, $V'(W) = V(W) + A$ with constant A) may be viewed as equivalent.

Let the proposition q apply only to a limited time interval t_q und a limited spatial region R_q. Then V(q) is to be understood as follows, where W_T is the true history:

(α) (a) If q is actually the case, then $V(q) = V(W_T)$.
(b) If q is false, then $V(q) = V(W_q)$, where W_q is the possible history of the world that would occur if q were always the case.

In (b) a counterfactual conditional is used. The explication of these is still controversial. For our purposes the following indications should suffice, though they would need to be made more precise. In the present context, we will use only counterfactuals in which the condition q is limited in the above way and moreover in which q is consistent with the totality PL of the actual physical laws (in the sense of §. . ., so not in the sense of the laws currently recognized by scientists). W_q is therefore the history of the world that meets the following conditions: [2]

(β) (a) W_q coincides with W_T over its entire range *before* the time interval t_q,
(b) as well as during the interval t_q *outside* the region R_q,
(c) within the space-time region $\{t_q, R_q\}$, W_q coincides as far as possible with W_T and diverges from W_T only as far as is necessary to make q true;
(d) *after* the interval t_q, W_q coincides with W_T in all space-time regions not affected causally by the previous q, while they diverge from W_T in the regions affected by q as determined by q in conjunction with the laws PL. [3a]

(γ) p is *better* than q with respect to the value function V $=_{Df}$ $V(p) - V(q) > 0$.
(δ) p is *good* with respect to the value function V $=_{Df}$ p is better than not-p. [3b]

Assuming that an agent X has a choice among the possible actions of a set A_X, we define:

(ε) The possible action a in A_X is an *optimum* with respect to the value function V $=_{Df}$ no action in A_X is better (in the sense of (γ)) than a with respect to V. [3c]

(22 February)

A person X at a given time has not just a *single* value function, but a great many of them, representing different value aspects. If X, following the dietary advice of his doctor, says "It is better for me to avoid a certain kind of food", he has a certain value function in mind, one that represents only health values, and only for himself. Other partial value aspects might be: his business profit, his aesthetic pleasure, his own well-being with respect to all aspects jointly, the well-being of a family, that of a large group, that of a nation, that of all humanity. But there is also a comprehensive value function of X that comprises all aspects, and in which the relative weight of each aspect in any possible overall situation finds expression—aspects that are sometimes in mutual conflict. Different things are meant by [the expression] “moral value judgement.” Perhaps it is best to use this term for the overall value judgement, in which the different aspects are included. [4]

The rational agent

(ζ) *Relative rationality* With respect to a value function V, a credibility function Cred, a body of evidence E and a set A of possible actions, an action a in A is *rational* $=_{\mathrm{Df}}$ for no action a' in A different from a is V(W) using Cred on the basis of E and a' preferred to V(W) on the basis of E and a. (The degree to which V(W) is preferred with respect to a certain body of evidence is the sum over all possible W of the products of V(W) with the credibility of W on the basis of the evidence in question; see § [25(II)].) [5]

There are certain standards on the basis of which a Cred-function can be criticized as irrational; these have been discussed elsewhere (Kemeny’s essay §[III]; and my §[26(IV)] in this reply). It is the task of inductive logic to arrive at such standards.

Are there also *standards of rationality for value functions?* The above-mentioned standards of inductive logic are not applicable here. The acceptance of a value function is completely independent of factual questions, for what the value function primarily evaluates is not particular actions or processes but rather entire possible histories of the world. Considerations about the consequences to be expected from an action do not come into the picture, for in a W all consequences are already included and given. [For instance, take the case where] the function V_1 values W_1 more highly than W_2, while the function V_2 does the reverse:

(a) $V_1(W_1) > V_1(W_2)$
(b) $V_2(W_1) < V_2(W_2)$.

Assume that the agent X_1 accepts V_1 and X_2 accepts V_2. Assuming that X_1 and X_2 discuss their value functions and, in particular, the descriptive results (a) and (b). In their discussion they will consider only the two histories W_1 und W_2. X_1 may have different evidence values than X_2 for each of these two histories; but that is irrelevant for the question of choosing between V_1 and V_2. This [6] question concerns only whether one values W_1 more highly than W_2 or vice versa; that has no bearing on the question whether W_1 will occur or has a higher probability [of occurring] than W_2.

Although all logic, including inductive logic, and factual knowledge are irrelevant, it nonetheless seems to me that there are other, purely valuational criteria by which to judge a value function as more or less rational than another. I am not going to attempt to set up fundamental standards for such judgements here. I only want to mention some considerations whose justification in such a judgement seems plausible and would likely be approved by most people, even if they diverge markedly in their valuations. First, it seems reasonable to require that a value function V(W) is derivable from general principles regarding the valuation of particular processes; specifically that the value of V(W) be an algebraic sum (or integral) of positive or negative values determined by some sort of principles governing certain very specific processes, while the remaining processes are irrelevant. (The relevant processes [7] arise e.g. from certain affective processes in humans, or from a more general kind of processes in beings that are animate or regarded as such; while the inorganic processes are of course irrelevant.) Then it should also be required that the principles have a general character, that they are expressible by mathematical functions of the relevant properties of the processes involved, specifically mathematical functions that are continuous and relatively smooth, rather than jumping up and down. These examples of requirements may be doubtful. I have not mentioned them to defend their validity, but only to indicate why I think that there are certain standards a value function must meet to be rational. The clarification of such standards I can't attempt here. But it seems clear that if such standards were worked out, they would only exclude as irrational certain value functions, and still admit an infinite set of different value functions that are extremely different from each other, and among them would be many that would be considered by most people, perhaps by all, as completely wrong and immoral. So the standards I speak of do not at all have the function of excluding "immorality" [8] or of distinguishing between value judgments that occur psychologically in controversies about moral or political questions. In the following I will speak of "the standards of rationality for value functions" as if they had already been arrived at. [9]

Now we define:

The behavior of an agent X is *perfectly rational* during a certain time period Δt when it meets the following conditions:

(η) (a) In *deductive thought*, which includes the whole of pure mathematics, he never makes any errors during Δt.
(b) During the period Δt he uses a rational method in his *inductive thought*; specifically, there is a *credibility* function Cred_X for him that meets the criteria of rationality.
(c) His behavior during the period Δt is governed (in the way to be described under (d)) by a *value function* V_X that meets all standards of rationality.
(d) Whenever X has a choice, at a time t within the period Δt, among different actions in a set $A_{X,t}$, and if at t his total evidence is $E_{X,t}$, then the action chosen by X has *relative rationality* (in the sense of ζ) with respect to V_X, Cred_X, $E_{X,t}$, and $A_{X,t}$. [10]

Assume X is perfectly rational at time t and chooses action a in A_X. Then it is nonetheless still possible for a *not to be an optimum* with respect to V_X. It could be that an action a' is better than a with respect to V_X, due to certain circumstances not known to X at the time of the action. It could even be that the objectively better, i.e. more successful action a' would not be rational for X. As emphasized elsewhere (§[26.IV]), rationality is not to be determined by success.

No one is ever perfectly rational in the sense just defined. "More rational", whether applied to different periods or to two possible behaviors of the same person in the same period, cannot very well be exactly defined. Roughly speaking, a behavior is more rational than another when it comes closer to perfectly rational behavior. But since deviations from perfectly rational behavior are possible in completely different ways, e.g. in the ways mentioned above (η) (a), (b), (c), (d), and within each of these once again in different ways, it is hardly possible to decide without an arbitrary convention under what conditions a deviation in one way should be considered equal to a deviation in another way.