The Defence of Utilitarianism in Early Rawls: A Study of Methodological Development

JUKKA MÄKINEN
Aalto University

MARJA-LIISA KAKKURI-KNUUTTILA
Aalto University

Rawls scholarship has not paid much attention to Rawls's early methodological writings so far, pretty much focusing on the *reflective equilibrium* (RE) which he is understood to have adopted in *A Theory of Justice*. Nelson Goodman's coherence-theoretical formulations concerning the justification of inductive logic in *Fact, Fiction and Forecast* have been suggested as the source of the RE. Following Rawls's methodological development in his early works, we shall challenge both these views. Our analysis reveals that the basic elements of RE can be located in his 'Two Concepts of Rules' essay. We shall further show that the origins of RE go all the way back to Aristotle's methods of ethics, as RE accords with the methodology entitled *saving the appearances* (SA) in recent Aristotle scholarship.

INTRODUCTION

The remarkable influence John Rawls's reflective equilibrium in his treatise *A Theory of Justice* (1971) has had on methodological discussions in various fields seems to have overshadowed his earlier methodological works.[1] Even though it is well known that Rawls started his career in ethics and political philosophy as a methodological thinker, his early methodological ideas and their relation to reflective equilibrium have not been sufficiently disclosed. In his dissertation, 'A Study in the Grounds of Ethical Knowledge: Considered with Reference to Judgments on the Moral Worth of Character' (1950), he states that to discover and validate principles supporting ethical decision-making is one of the main problems of philosophy.[2] Rawls's rhetorical strategy

[1] J. Rawls, *A Theory of Justice* (Oxford, 1971).

[2] J. Rawls, 'A Study in the Grounds of Ethical Knowledge: Considered with Reference to Judgments on the Moral Worth of Character' (PhD dissertation, Princeton University, 1950), p. 1.

Utilitas Vol. 25, No. 1, March 2013
doi:10.1017/S0953820812000222

indeed seems quite clever since, to carry out this task, he developed a scientistically oriented methodology for ethics, thus attacking the positivist rejection of ethical theorizing with its own weapons. The methodology was inspired by the newly discovered inductive logic, which explains our use of the title *inductive logic* (IL for short) methodology for it.[3] Rawls then repeats the chief features of IL almost unchanged in his 'Outline of a Decision Procedure for Ethics'.[4] In another early, much-studied contribution, 'Two Concepts of Rules',[5] he sets out to illustrate the use of IL without repeating its details, leaving the reader with a reference to their exposition in the 'Outline'.

In each of these three early writings, Rawls offers several examples intended to illustrate how the IL methodology works in practice. Studying the argument structure of these illustrations one may, however, see substantial methodological discrepancies between the proposed theory and actual methodological practice. In spite of the great hopes Rawls invested in IL, his intended illustrations of it simply fail to fit his scientistic methodology. Therefore the chief aim of this article is to trace Rawls's methodological development in these three early writings by investigating his methodological practice in each.

While the IL methodology is mainly inductivist, as we shall show, the argument structure of the intended illustrations both in the dissertation and in the 'Outline' essay fits the hypothetico-deductive model better. Our most exciting finding, however, concerns the 'Two Concepts' essay, in which Rawls offers solutions to two theoretical problems, one concerning the justification of punishment and the other that of promise. One may note here a significant change as compared with examples in the two earlier writings, as neither solution accords with the IL methodology or the hypothetico-deductive model. To put it in standard epistemological terms, the change involves a move from a strong foundationalist emphasis to clear coherence-theoretical thinking. To clarify this change further, we shall show that a revealing description of Rawls's argument practice in 'Two Concepts' can be offered with the help of Aristotle's methodology in ethics, often called *saving the appearances* (SA for short), which has only recently been brought to light in Aristotle scholarship. The designation 'saving the appearances' has been adopted from the title of G. E. L. Owen's trail-blazing 1961 essay 'Tithenai ta phainomena'.[6] Since the Aristotelian

[3] Rawls, 'A Study', pp. 68–9; J. Rawls, 'Outline of a Decision Procedure for Ethics', *Philosophical Review* 60 (1951), pp. 177–97, at 178, 189.

[4] Rawls, 'Outline'.

[5] J. Rawls, 'Two Concepts of Rules', *Philosophical Review* 64 (1955), pp. 3–32.

[6] G. E. L. Owen, 'Tithenai ta phainomena', *Logic, Science and Dialectic: Collected Papers in Greek Philosophy*, ed. M. C. Nussbaum (London, 1987), pp. 239–51; G. E. R.

SA can be shown to yield a *reflective equilibrium* (RE for short) of prior beliefs, this indicates that Rawls's now famous reflective equilibrium approach was already in use in the 'Two Concepts'.

So far, Rawls scholarship has not paid much attention to his early methodological development, pretty much unanimously focusing on reflective equilibrium, which is understood as having been adopted as late as his *A Theory of Justice*. Even though the history of the term 'equilibrium' goes back at least to Leibniz,[7] Norman Daniels cites Nelson Goodman's coherence-theoretical formulations concerning the justification of inductive logic in *Fact, Fiction and Forecast* (1955) as the source of Rawls's RE.[8] As already alluded to, our developmental analysis leads us to challenge both these chronologies. The conception that reflective equilibrium appears in *A Theory of Justice* for the first time is undermined by our analysis of Rawls's defence of utilitarianism in his justification of punishment in 'Two Concepts', which reveals that the basic elements of RE can be located in this 1955 essay. Since the roots of RE go back to Aristotle, as we shall show, and Rawls himself cites a chief methodological passage from Aristotle's *Nicomachean Ethics* VII.1, Goodman is hardly the main source for the Rawlsian RE. Goodman's book appeared the same year as the 'Two Concepts' essay and, furthermore, inductive logic forms an important reference point for Rawls when arguing for the reasonableness of his inductivist IL methodology in the dissertation.[9]

To get a better grasp of the argument of this article, it is worth taking a look at how the reflective equilibrium is understood in the Rawls scholarship. Even though some of Rawls's characterizations of RE remain ambiguous, he often emphasized that, instead of describing a research methodology, RE characterizes first of all the *state* achieved as the end result of research.[10] However, the title 'reflective equilibrium' has occasionally been adopted not only for a coherence-theoretical *justification* of an ethical or political theory consisting of

Lloyd, *Aristotle: The Growth and Structure of his Thought* (Cambridge, 1968); M. C. Nussbaum, *The Fragility of Goodness: Luck and Ethics in Greek Tragedy and Philosophy* (Cambridge, 1986); C. Witt, 'Dialectic Motion and Perception: De Anima Book I', *Essays on Aristotle's De Anima*, ed. M. C. Nussbaum and A. O. Rorty (Oxford, 1992), pp. 169–83. Cf. W. D. Ross, *Aristotle* (London, 1971/1923) on methodology of ethics.

[7] J. Rawls, *Lectures on the History of Moral Philosophy*, ed. B. Herman (Cambridge, MA, 2000), pp. 105–40, at 136–7.

[8] Norman, Daniels, 'Reflective Equilibrium', *Stanford Encyclopedia of Philosophy* <www.plato.stanford.edu/entries/reflective-equilibrium/> (2011).

[9] Rawls, 'A Study', pp. 68–9.

[10] J. Rawls, *A Theory of Justice*, pp. 46–7. Cf. J. Rawls, 'The Independence of Moral Theory', ed. S. Freeman, *John Rawls: Collected Papers* (Cambridge, MA, 1975/1999), pp. 286–302, at 288–301.

a reflective equilibrium of preceding views on the matter,[11] but also for a full-blown methodology which is supposed to end in achieving a reflective equilibrium. For instance, Norman Daniels describes RE as a methodology which embraces both research heuristics, i.e. the process of discovery, as well as justification of the end result:

> The method of reflective equilibrium consists in working back and forth among our considered judgments (some say our 'intuitions') about particular instances or cases, the principles or rules that we believe govern them, and the theoretical considerations that we believe bear on accepting these considered judgments, principles, or rules, revising any of these elements wherever necessary in order to achieve an acceptable coherence among them. The method succeeds and we achieve reflective equilibrium when we arrive at an acceptable coherence among these beliefs. An acceptable coherence requires that our beliefs not only be consistent with each other (a weak requirement), but that some of these beliefs provide support or provide a best explanation for others. Moreover, in the process we may not only modify prior beliefs but add new beliefs as well.[12]

Daniels's description of RE includes the following three phases:

> (D 1) Collecting considered judgements of particular instances or cases, principles or rules we believe govern the considered judgements, and theoretical considerations that we believe bear on accepting both the considered judgements and the principles.
>
> (D 2) The research process consists of working back and forth among the aforesaid beliefs by modifying them and adding new beliefs when needed.
>
> (D 3) The aim of the research process is to construct an acceptable, coherent set of the given beliefs.

His account of an acceptable coherent set of beliefs accords with the standard explication of coherence theory of justification, stating that, in addition to consistency, the beliefs in the end result support some of the other beliefs or provide a best explanation for them.[13]

Daniels's RE description suits a research process that begins with a set of mutually conflicting prior beliefs. This is exactly the case in 'Two Concepts', where Rawls faces two conflicting views concerning

[11] W. Van der Burg and T. Van Willigenburg, 'Introduction', *Reflective Equilibrium Essays in Honour of Robert Heeger*, ed. W. Van der Burg and T. Van Willigenburg (Dordrecht, 1998), pp. 1–25, at 1–2); N. Daniels, *Justice and Justification: Reflective Equilibrium in Theory and Practice* (Cambridge, 1996), pp. 1–2, 6; T. M. Scanlon, 'Rawls on Justification', *The Cambridge Companion to Rawls*, ed. S. Freeman (Cambridge, 2003), pp. 139–67, at 140–1. For a foundationalist interpretation of RE, see M. R. DePaul, *Balance and Refinement: Beyond Coherence Methods of Moral Inquiry* (London, 1993).

[12] Daniels, 'Reflective Equilibrium'.

[13] For coherence theory, see for instance M. Lammenranta, 'Theories of Justification', *Handbook of Epistemology*, ed. I. Niiniluoto, M. Sintonen and J. Wolenski (Dordrecht, 2004), pp. 479–82; M. Williams, *Problems of Knowledge: A Critical Introduction to Epistemology* (Oxford, 2001), pp. 117–27.

justification of punishment. This marks a clear contrast to his two earlier works, the dissertation and the 'Outline' essay, where both the IL methodology and the examples offered to illustrate it deal primarily with mutually consistent views which impose no particular challenge to modify prior beliefs when constructing a theory to cover them. We shall show that Rawls's argument structure in dealing with the justification of punishment issue fits with Daniels's description of RE, since Rawls's solution to the conflict between utilitarian and retributive justifications of punishment not only saves utilitarianism, but is a synthesis of both conceptions suitably modified. The methodologically important point here is that the Aristotelian saving the appearances offers a more detailed reading of the solution than Daniels's 'working back and forth' account. Our aim is thus to show that Aristotle's saving the appearances is a research methodology with heuristics for discovery and justification that ends up in a reflective equilibrium of prior views.

So far, only a few scholars have paid attention to the similarities between Rawls's and Aristotle's methodologies. A brief note on possible methodological parallels between *A Theory of Justice* and Aristotle's *Nicomachean Ethics* can be found in Jonathan Barnes's 'Aristotle and the Methods of Ethics', some comments in Sherwin Klein's 'The Value of *Endoxa* in Ethical Argument', and a more expanded analysis in Martha Nussbaum's 'Equilibrium: Scepticism and Immersion in Political Deliberation'.[14] Comparing Rawls's ethical argument with that of Aristotle may not be so far-fetched as it might seem to start with, since Rawls himself cites the most famous passage relevant to the Aristotelian SA in his dissertation, refers to a 'time-honoured' methodological device in 'Two Concepts', and mentions Aristotle's ethics methodology again in *A Theory of Justice*.[15] The fact that the details of the SA methodology have been explicated only recently explains why Rawls says so little about the connection between his and Aristotle's methodology in practical philosophy. Here we shall leave aside the important question of how Rawls came to adopt the SA approach.

In developing our argument we shall proceed as follows. The next section will be devoted to Rawls's understanding of his inductive logic methodology in ethical decision-making. In that connection, we shall also discuss some intended illustrations of IL from both the dissertation

[14] Jonathan Barnes, 'Aristotle and the Methods of Ethics', *Revue Internationale de Philosophie* 33–4 (1980), pp. 490–511; Sherwin Klein, 'The Value of *Endoxa* in Ethical Argument,' *History of Philosophy Quarterly* 9 (1992), pp. 141–57; M. C. Nussbaum, 'Equilibrium: Scepticism and Immersion in Political Deliberation', *Ancient Scepticism and the Sceptical Tradition*, ed. J. Sihvola (Helsinki, 2000), pp. 171–97.

[15] Rawls, 'A Study', p. 345; Rawls, 'Two Concepts', p. 7; Rawls, *Theory*, p. 51 n. 26.

and the 'Outline' to exemplify their poor fit with the characterization of IL. With the help of an example from Aristotle's *Nicomachean Ethics*, we shall, however, show the relevance of IL to ethical theorizing. The SA methodology will then be presented, and that section will end with a comparison of IL and SA. We shall then offer our reading of Rawls's JP argument in his 'Two Concepts of Rules' as an application of SA, and indicate how it complements Daniels's description of RE. To conclude, we shall offer a summary of our findings.

RAWLS'S INDUCTIVE LOGIC METHODOLOGY

Rawls's challenge to positivism

In his dissertation, Rawls offers a methodology to discover and validate ethical principles which will function as 'adequate and justifiable canons for the solution of moral conflicts', as well as 'serve as a means for the reform and improvement of common morality'.[16] Developing such a normative approach in ethics was no easy task in the prevailing academic atmosphere, however. The dominant positivist philosophy was far from friendly towards such normative projects, accepting merely linguistic analysis of ethical and political terminology. Rawls explains the popularity of the positivist position on ethics by the combination of two factors: the lack of a suitable methodology in ethics, and ethical relativism supported by the interpretations of current anthropological research.[17]

Thus the first step Rawls sees fit to adopt in his dissertation is to refute the authoritarian and positivist emotivist stands on ethics, both of which regard moral reasoning as futile. His decisive move is to develop a new methodological conception of ethics designed not to be easily rejected by his opponents.[18] Rawls's rhetorical strategy consists of two apparently irrefutable moves, the first arguing for the necessity of rational moral reasoning, and the second for its possibility, the latter comprising the main project of the dissertation. The necessity of well-established ethical principles for Rawls follows from the democratic conception of government, according to which laws founded on reasoned public discussion constitute the primary source of political authority. This implies that the modes of rational foundation of ethical principles should form a part of democratic theory, as well as ethical philosophy.[19]

[16] Rawls, 'A Study', pp. 87, 95. Rawls's IL is not merely a method, but a full methodology, since it includes an epistemology as well.

[17] Rawls, 'A Study', pp. 12–15.

[18] Rawls, 'A Study', p. 7.

[19] Rawls, 'A Study', pp. 7–8.

Since Rawls's strategy in the dissertation is to construct a scientistic methodology for ethics in simulation of the prevailing understanding of the natural sciences along the lines of the new inductive logic, this seems to undermine the claim that Goodman's comments on inductive logic in his 1955 work could have been a major methodological inspiration for Rawls.[20] Speaking about 'ordinary methods', Rawls removes the positivist demarcation between science and non-science by expanding the positivist unity of science principle to include ethics as well. Analogous to natural science, the role of ethical principles is to capture the invariant in actual ethical decisions, and in the case of several sets of principles, the criteria of simplicity and elegance are to guide the choice, just as is often suggested for natural science.[21] The only major modification required to make the IL methodology applicable to ethics, for Rawls, involves a new reading of 'data'. Instead of observations obtained by scientists, ethical data consist of *rational judgements* (*considered judgements*, following Rawls's later terminology) on ethical matters made by *reasonable men* (*competent judges*, also following Rawls's later terminology).[22] The criteria for the competent judges, as well as those for the considered judgements, thus form the main issues in his endeavour to establish an ethical methodology acceptable to the positivistically minded philosophers.

In our explication of Rawls's scientistically oriented IL methodology, we shall rely both on his dissertation 'Study in the Grounds of Ethical Knowledge: Considered with Reference to Judgments on the Moral of Worth of Character', and on the essay 'Outline of a Decision Procedure for Ethics'. As suggested by the opening words of the 'Outline' essay, Rawls aims to tackle two questions simultaneously:

> Does there exist a reasonable decision procedure which is sufficiently strong, at least in some cases, to determine the manner in which competing interests should be adjudicated, and, in instances of conflict, one interest given preference over another; and, further, can the existence of this procedure, as well as its reasonableness, be established by rational methodologies of inquiry?[23]

The first challenge is to find a reasonable decision procedure in situations of conflict of interest in ethical matters, and the second is to establish the reasonableness of the methodology itself. The basically

[20] Rawls, 'A Study', pp. 68–9; Rawls, 'Outline', pp. 178, 189.

[21] Rawls, 'Outline', p. 186.

[22] We shall not explore to what extent Rawls's IL owes to Sir David Ross's conception of the method of ethics in W. D. Ross, *The Right and the Good* (Oxford, 1930). On this see Sherwin Klein, 'The Value of *Endoxa* in Ethical Argument', *History of Philosophy Quarterly* 9 (1992), pp. 141–57, at 148.

[23] Rawls, 'Outline', p. 177.

foundationalist nature of Rawls's methodology becomes clear from the following summary of its three main steps:

(IL 1) Collecting the considered judgements on particular moral conflict situations.
(IL 2) Explicating ethical principles on the basis of these considered judgements.
(IL 3) Justifying these principles by showing that they comprehensively cover the given as well as new considered judgements on the matter.[24]

The challenge of showing the reasonableness of the IL methodology itself will be met by demonstrating that it serves its purpose of supporting ethical decision-making by yielding ethical principles which cover good ethical decisions already made, and to be made in situations of conflict of interest. Next we shall present Rawls's IL methodology in some detail, and briefly assess the extent to which these two goals are achieved in the dissertation and the 'Outline'.

Considered judgements and competent judges

The first phase in the Rawlsian IL is to gather the relevant 'data' for constructing the ethical principles (IL 1). Rawls's notion of ethical 'data' clearly differs from simple observations typically construed as the knowledge basis of science by empiricists, since the 'data' is said to consist of actual decisions made in situations of conflict of interest.[25] Such decisions are characterized as follows: 'since A, B, C, . . ., and M, N, O, . . . are the facts of the case and the interests in conflict, M is to be given preference over N, O, . . .'.[26] The decision suggests simply, without compromise, whose preferences are to be preferred over those of others. To avoid circularity, choice is to be made intuitively without relying on any ethical principles, and one is to feel certain about it.[27]

To form a sound basis for ethical theorizing, the decision needs to be made under favourable conditions by a normal, intelligent person meeting certain further requirements to make him a 'competent judge'. Rawls characterizes the competent judge as one who is knowledgeable about the relevant features of human action in general, such as the likely consequences of an action. He should also be knowledgeable about human interests in a sympathetic manner, and capable of using his imagination in cases where he lacks experience. In order to make a fair and informed choice, the judge has to inquire into all the relevant facts of the case, and allow each party a fair opportunity to state his case. In

[24] Rawls, 'Outline', pp. 178–90.

[25] Felt meanings do not function as a basis for testing moral principles, only actual judgements. Rawls, 'A Study', pp. 75–7; Rawls, 'Outline', p. 185.

[26] Rawls, 'Outline', pp. 177–8, 186; cf. Rawls, 'A Study', p. 57.

[27] Rawls, 'Outline', pp. 45–6, 57–9.

addition to possessing at least normal intelligence, a competent judge must be willing to use the criteria of inductive logic to determine with an open mind what to believe and what not. To guarantee neutrality, he should be aware of the possible influences of his own emotional, intellectual and moral predilections.[28]

Instead of being one of the parties to the conflict, the competent judge is to be an impartial third party with no gain or loss to expect from his decision, whatever it may be. A considered judgement is further required to be stable in the sense that one can find several competent judges who have rendered an identical decision in a similar situation at other times and places. In other words, the judgements relevant to the construction of ethical principles are such that they 'are made from day to day on the moral issues which continually arise'.[29]

These features of the competent judge can be assessed independently of the kind of judgements he offers, which guarantees a kind of objectivity for the evidence base of the moral principles in a non-circular manner.[30] Rawls calls this conception of moral data 'logical' or 'methodological physicalism', by which he means 'a principle which embodies an essential rule of scientific methodology, namely, the insistence that the theories and principles of a science be established or refuted wherever possible by objective data which can be checked by the community of investigators, together or individually'.[31]

Explication of ethical principles

Having gathered a sufficient set of considered judgements about a number of different conflict situations, the next step is to formulate, *explicate* in Rawls's terminology, general moral principles to capture what is invariant in the considered judgements (IL 2).[32] Even though Rawls speaks about explication as a *heuristic device*, he says nothing about the intermediate steps to be taken between considered judgements and principles.[33] The examples meant to exemplify the working of IL in both the dissertation and the 'Outlines' offer no help in this respect, since they merely illustrate the justification of ready principles. This accords with his statement that 'the norms of inductive logic are for the purpose of evaluating the truth of theories once they

[28] Rawls, 'A Study', pp. 37–8; Rawls, 'Outline', pp. 178–81.

[29] Rawls, 'Outline', p. 183; see also Rawls, 'A Study', pp. 49–52.

[30] Rawls, 'A Study', pp. 32, 44; Rawls, 'Outline', pp. 181–3. Rawls rejects the choice of judges on the basis of ideology, social class, institutional group, and race (Rawls, 'A Study', p. 32; Rawls, 'Outline', p. 181).

[31] Rawls, 'A Study', p. 78.

[32] Rawls, 'A Study', pp. 68–70.

[33] Rawls, 'A Study', pp. 68–9.

are formulated. They are not rules for discovery, but canons of proof.'[34] This is, however, not in harmony with several claims stating that both discovery and justification are equally crucial elements of IL.[35] In Rawls's examples intended to illustrate his IL, we shall see that its second phase, explication of the principles on the basis of considered judgements (IL 2), plays a minor role after all.

Justification of ethical principles

The task of the moral principles with respect to considered judgements resembles the covering law model of explanation, although Rawls points out that the principles are not intended to offer the causes of the considered judgements.[36] Likewise, an explication is required to be *comprehensive* so that it covers 'all considered judgements, and it is expected to do this with the greatest possible simplicity and elegance'.[37] This is the chief form of justifying an explication: 'Like any theory, an explication is tested by the criterion of comprehensiveness.'[38] The set of principles, if it exists, will be useful in practical life by helping to assess what interests to prefer in situations of ethical conflict.[39] This means that applying the principles to new cases would yield judgements identical to those made intuitively by the competent judges.[40]

Rawls's strong emphasis on comprehensiveness as the criterion of acceptable ethical principles leads us to classify his IL as basically a foundationalist epistemology with the considered judgements as the immediately justified, logically consistent and irrefutable basis, and the moral principles as the mediately justified beliefs.[41] The Rawlsian IL is not, however, a purely foundationalist epistemology, since he points out three further situations one may face when testing the success of an explication. The first involves conflicting judgements among different judges, the second an anomaly (a judgement apparently not covered by the principles) and the third a judgement in conflict with some principle. The principles are shown to be reasonable where a solution

[34] Rawls, 'A Study', p. 68.

[35] Rawls, 'A Study', pp. 1, 16, 50, 61, 103.

[36] Rawls, 'A Study', pp. 79–81; Rawls, 'Outline', p. 185.

[37] Rawls, 'Outline', p. 186. The situation in which one should choose between several equally adequate theories forms no objection to his approach since, according to Rawls, the actual instances of several adequate and comprehensive theories are non-existent. Rawls, 'A Study', pp. 81–4.

[38] Rawls, 'A Study', p. 86.

[39] Rawls, 'A Study', p. 87; Rawls, 'Outline', p. 186.

[40] Rawls, 'Outline', p. 186.

[41] For foundationalist epistemology, see for instance, M. Lammenranta, 'Theories of Justification', *Handbook of Epistemology*, ed. I. Niiniluoto, M. Sintonen and J. Wolenski (Dordrecht, 2004), pp. 467–97, at 473–6; M. Williams, *Problems of Knowledge: A Critical Introduction to Epistemology* (Oxford, 2001), pp. 81–5.

to such problems can be brought about 'which, after criticism and discussion, seems to be acceptable to all, or nearly all, competent judges, and to conform to their intuitive notion of a reasonable decision'.[42] The third case allows a fourth form of justification, since, when a principle survives the conflict with a considered judgement and the latter is modified, it is desirable to offer the reason for the mistake in the considered judgement.[43]

The capacity to resolve anomalies and conflicts among considered judgements, and conflicts between a principle and a considered judgement is, for Rawls, a powerful guarantee of the reasonableness of the principles established. Such forms of justification constitute non-foundationalist, coherence-theoretical elements in his IL, since not all considered judgements are regarded as infallible while some may be modified by the principles which again are supported by the modified judgements. The foundationalist element, in contrast, consists of the idea that the considered judgements form a firm knowledge basis on which to explicate general ethical principles, which are supported inductively by these same considered judgements as well as new ones. The coherentist element involves the possibility of correcting a subclass of considered judgements on the basis of the principles explicated, thus allowing a relation of mutual support between the modified considered judgements and the ethical principles explicated.[44] As the emphasis on comprehensiveness clearly indicates, the foundationalist element is for Rawls the main one while the coherentist perspective plays a minor role in IL. In practical applications coherence considerations could, in principle, turn out central in case the considered judgements were mostly inconsistent. As we shall show there is a drastic change in this respect in Rawls's examples in the dissertation and 'Outline' essay as compared with the 'Two Concepts' essay.

Rawls offers four distinct arguments for the reasonableness of the IL methodology, i.e. to support the view that IL is a reasonable way to establish principles to support ethical decision-making in situations of conflict of interest. Because of its strong foundationalist emphasis

[42] Rawls, 'Outline', p. 188; see also Rawls, 'A Study', p. 92.

[43] Rawls, 'Outline', p. 189. A considered judgement may be modified by deciding on another preference order because of seeing that some relevant fact of the situation has previously been ignored (Rawls, 'Outline'). Another modification is to reject a judge as competent, allowed by Rawls's remark that the tests for selecting the judges are bound to remain vague (Rawls, 'A Study', p. 41; Rawls, 'Outline', p. 180).

[44] For coherence theory, see for instance M. Lammenranta, 'Theories of Justification', *Handbook of Epistemology*, ed. I. Niiniluoto, M. Sintonen and J. Wolenski (Dordrecht, 2004), pp. 467–97, at 479–82; M. Williams, *Problems of Knowledge: A Critical Introduction to Epistemology* (Oxford, 2001), pp. 117–27. Such coherentist elements do not, according to Rawls, involve a move away from the scientistic model, but are analogous to justification in natural science (Rawls, 'A Study', pp. 93–4; Rawls, 'Outline', p. 184).

IL presupposes a privileged position among moral judgements for the considered judgements. This is what they have, according to Rawls, since the considered judgements are made in favourable conditions to minimize the influence of personal predilections, and their neutrality is further guaranteed by the requirement that several persons have made the same judgement.[45] Rawls's second argument for the reasonableness of IL is that the way of justifying IL resembles the justification of inductive logic, the soundness of which is established by showing that it yields the modes of reasoning applied by scientists.[46] Third, such an IL methodology Rawls claims to be superior to various traditional forms of justifying ethical principles, such as reliance on ethical intuition, the authority of reason, divine revelation or analysing the meaning of ethical terms.[47] The fourth argument for the reasonableness of IL is supposed to be given by demonstrating that the methodology actually works in practice. We shall next take a look at some typical examples Rawls offers as illustrations of the IL methodology in order to assess how far they accord with his portrayal of the three stages of IL.

Assessment of intended illustrations of IL in the dissertation and 'Outline'

In Part II of the dissertation, Rawls claims to offer a host of examples as illustration of the IL methodology. One of his aims is said to '*explicate* the rational judgements of reasonable men so far as they are applied to the moral character of an agent'.[48] A closer look at the examples reveals, however, that instead of taking a set of considered judgements in situations of conflict of interests as starting points for explication, as suggested by the first phase of IL (IL 1), the treatment begins with ready principles, presuming that these are already available without a process of discovery. This implies that the intended illustrations begin with the third and final step (IL 3), and fail to exemplify the first two steps (IL 1) and (IL 2).

To elucidate, one of Rawls's principles concerning moral character reads as follows: 'The character of an agent who merely contemplates the doing of an evil action, but does not do it, is not to be judged as bad as the character of an agent who not only contemplates it, but does it.' Instead of relying on considered judgements as required

[45] Rawls, 'A Study', pp. 61–2; Rawls, 'Outline', pp. 187–8.

[46] Rawls, 'Outline', pp. 177–8, 188–90 and 195–6.

[47] Rawls, 'Outline', pp. 184 and 197. Rawls strongly rejects metaphysical approaches to ethical issues, claiming that the existence of ideal values, causes of moral judgements and the universality of moral codes has no bearing on the objectivity or subjectivity of moral knowledge (Rawls, 'A Study', p. 106; Rawls, 'Outline', p. 177).

[48] Rawls, 'A Study', p. 103.

in (IL 3), the justification of this and the other four principles dealt with relies partly on two principles of practical reason pertaining to the goal-directness of intentional action,[49] and partly on some general common-sense conceptions,[50] as well as teachings from the New Testament.[51]

Rawls's discussion of another principle that concerns one who merely contemplates the doing of a right action, but does not do it, ends likewise with the remark that it is justified by the principles of practical reason.[52] The principles of practical reason utilized as evidence are:

> (i) that it is reasonable to adopt appropriate means to appropriate ends; (ii) that the reasonable and appropriate ends are those activities which are comprehensively satisfactory for the individual person and also inclusively harmonious with like and other activities of the members of the community in which he lives.[53]

The treatment of the principles of justice in the 'Outline' essay have a similar structure. At the outset we are given seven principles with several sub-principles, such as, 'each claim in a set of conflicting claims shall be evaluated by the same principles', and 'given a group of competing claims, as many as possible shall be satisfied, so far as the satisfaction of them is consistent with other principles'. To illustrate their justification, a few of them are shown to be in harmony with some general conceptions, such as freedom of speech and thought, and the rejection of the institution of inquisition.[54] Here again the process of discovery, i.e. steps (IL 1) and (IL 2), are left out and the evidence applied in the justification consists of general principles instead of decisions resembling the account of the considered judgements.

In summary of our assessment of Rawls's intended illustrations of IL both in the dissertation and the 'Outlines' we claim that he fails to demonstrate that the IL methodology works in actual research practice to guarantee sound principles for ethical decision-making. Our analysis of the illustrations reveals a lack of the phase of collecting the considered judgements (IL 1), as well as the discovery phase (IL 2), even though both discovery and justification are purported to be equally crucial elements of IL. Instead of presenting all three steps of IL, Rawls moves directly to ready ethical principles, merely exemplifying their justification (IL 3). Thus, instead of illustrating the inductivist methodology, Rawls's examples both in the dissertation and in the

[49] Rawls, 'A Study', pp. 105–6.
[50] Rawls, 'A Study', p. 122.
[51] Rawls, 'A Study', pp. 126–30, 132–4.
[52] Rawls, 'A Study', p. 139.
[53] Rawls, 'A Study', pp. 105–6.
[54] Rawls, 'Outline', pp. 194–5.

'Outline' essay can be read as illustrations of the hypothetico-deductive model. His hypothetico-deductive model is not, however, the standard empiricist one, since the justification of general principles relies on other general principles and general common-sense conceptions. The considered judgements in the sense of intuitive decisions in situations of ethical conflict of interest do not have a role in the justification process, either. No matter how reasonable IL is in theory, Rawls's main argument for its reasonableness collapses, as he fails to demonstrate that IL works in practice to generate ethical principles to guide decision-making in situations of conflict of interest.

To show that this should not be read as implying that an inductivist approach along the lines of the Rawlsian IL has no relevance to ethical theorizing, we shall briefly consider Aristotle's argument for the relevance of the mean in *Nicomachean Ethics* II.6 that includes an inductive generalization. The first step in the argument relies on expert opinion, e.g. trainer's decisions concerning the amount of nourishment given to athletes, runners and wrestlers, which should neither be too little, nor too much, but right for the sportsman in question. This evidence, somewhat comparable to Rawls's considered judgements made by competent judges, supports the following generalization: 'a master of any art avoids excess and defect, but seeks the intermediate and chooses the this-intermediate not in the object but relatively to us'.[55] Next, Aristotle argues that, likewise, a virtuous person strikes a mean in his actions which is not an arithmetical mean, but one based on contextual judgement.[56] This principle is thus gained by generalizing on the basis of consistent conceptions, a move in harmony with the IL methodology.

We may conclude this section on Rawls's inductive logic methodology by suggesting that even though Rawls's examples fail to display how his IL functions in practice, there is no reason to reject totally the inductivist approach in ethical theorizing. Rawls's own examples indicate that the inductivist approach should be complemented by the hypothetico-deductive model. The purpose of the rest of the article is to demonstrate that ethical theorizing is in need of stronger methodological devices than the inductivist and the hypothetico-deductive approaches taken together. The basic reason for this is fairly obvious: the 'data' the ethical theoretician is to work with seldom forms a consistent set, but often includes mutually inconsistent beliefs. The Aristotelian saving the appearances and the Rawlsian reflective equilibrium prove their strength exactly in cases like that.

[55] *Nicomachean Ethics* II.6, 1106b5–7.
[56] *Nicomachean Ethics* II.6.

ARISTOTLE'S SAVING THE APPEARANCES METHODOLOGY

We shall next show that the Aristotelian saving the appearances methodology offers a more detailed description of the research process leading to reflective equilibrium than Daniels's three-step characterization. This can be explained historically by Aristotle's ambition to build on Plato's achievements in experimenting with the possibilities of different argument strategies, which led Aristotle to develop a heuristics for creating a coherent synthesis out of prior beliefs on the basis of their critical scrutiny.[57] The Aristotelian research model thus helps us to explicate in more detail what Daniels's second step (D 2), i.e. 'working back and forth' among the prior beliefs, involves. As several studies indicate, Aristotle's own methodological practice neatly accords with his own brief methodological comments.[58] We need to keep in mind, however, that Aristotle's SA is not offered as a full heuristic of theory construction, but is confined to solving a particular theoretical problem at a time. At the end of this section we shall argue that IL and SA are two distinct methodological conceptions.

The four phases of SA

The following quotation from *Nicomachean Ethics* VII.1 is Aristotle's famous passage describing his understanding of the research process when facing problems concerning how to specify particular theoretical notions in a systematic step-by-step approach. Even though the passage is also quoted by Rawls in his dissertation,[59] he makes no further use of it, apparently because its peculiar details were brought to light only somewhat later. Here the problems concern weakness of will (*akrasia*).

> Here, as in all other cases, we must set down the appearances (*phainomena*) and, first working through the puzzles, in this way go on to show, if possible, the truth of all the reputable opinions (*endoxa*) about these affections or, if this is not possible, of the greater number and the most authoritative. For if the difficulties are resolved and the reputable opinions (*endoxa*) are left in place, we will have done enough showing.[60]

[57] M.-L. Kakkuri-Knuuttila, 'The Role of the Answerer in Plato and Aristotle', *Dialectic and Dialogue: The Development of Dialectic from Plato to Aristotle*, ed. J. Fink (Cambridge, in press).

[58] G. E. R. Lloyd, *Aristotle* (Cambridge, 1968); M. C. Nussbaum, *The Fragility of Goodness* (Cambridge, 1986); C. Witt, *Dialectic: Essays on Aristotle's De Anima*, ed. M. C. Nussbaum and A. O. Rorty (Oxford, 1992), pp. 169–83.

[59] Rawls, 'A Study', p. 345.

[60] *Nicomachean Ethics* VII 1, 1145b2–7. Translations of passages from the *Nicomachean Ethics* are from the Barnes edition.

The passage reveals four stages in a philosophical research process:

(SA 1) Collecting the appearances (*phainomena*) (SA 1).
(SA 2) Critical assessment of the given appearances.
(SA 3) Constructing the solutions to the problems found in phase (SA 2).
(SA 4) Justification of the solutions.

Here again the aim of the research process is to construct an acceptable, coherent set of the given beliefs as in Daniels's description of the process leading to reflective equilibrium. In order to demonstrate that and how phases (SA 2) and (SA 3) may offer heuristic advice for working back and forth through the given appearances and end up in a reflective equilibrium we need to take a detailed look at each phase of Aristotle's saving the appearances.

Collecting ethical 'data' in SA

First we need to show that Aristotle's 'appearances' cover Daniels's starting points of ethical research, i.e. include 'considered judgments of particular instances or cases …, principles or rules we believe govern the considered judgments, and theoretical considerations that we believe bear on accepting both the considered judgments and the principles'. The term 'appearance' expresses how matters appear to us, and it includes not only how they do so in observation and experience, or in intuitive judgements, but also in theoretical reasoning.[61] Saving appearances relevant to ethical issues thus means saving current views on ethical matters at various levels of abstractness.

The term 'appearance' is ambiguous in an interesting way that reveals a major difference between SA and the foundationalist emphasis in Rawls's IL. Things may appear to us in a misleading way, but also reveal what is truly the case, as exemplified by over-generalizations reflecting the particular social position of a group of people.[62] For instance, the highest goal of human life appears to be different to people in different circumstances, like a politician, one seeking wealth or pleasure, or a healthy and a sick person.[63]

Another epistemic notion characterizing the basis and criterion of philosophical inquiry in Aristotle is 'reputable opinions' (*endoxa*),

[61] This clarification of the notion of 'appearance' in Aristotle explains the major difference between the Aristotelian and the instrumentalist conceptions of saving the appearances. In the instrumentalist reading of astronomy, 'appearance' designates observations, and thus saving the appearances in astronomy means the systematization of *observations* carried out with the help of mathematical models: see P. Duhem, *To Save the Phenomena: An Essay on the Idea of Physical Theory from Plato to Galilei* (Chicago, 1908/1969).

[62] Barnes, 'Aristotle', p. 491 n. 1.

[63] *Nicomachean Ethics* I.5.

meaning the opinions of reputable men (*endoxoi*).[64] Excluding children and madmen, Aristotle's classification of reputable opinions in his textbook on dialectic, the *Topics*, embraces the opinions of all, the majority, the wise or the most authoritative of the wise, as well as experts in various fields.[65] The term 'reputable' expresses the social aspect of the knowledge basis, since it is social recognition that lends credibility to the beliefs of reputable people, and makes them worthy of serious consideration. Even though their connotations differ, in ethics the extensions of the terms 'appearances' and 'reputable opinions' seem to coincide.[66]

This shows that the starting points for ethical theorizing in the Aristotelian SA include considered judgements with their fairly strict requirements imposed by the Rawlsian IL, but they also embrace the other elements Daniels includes in the starting points for building a reflective equilibrium, i.e. 'principles or rules we believe govern the considered judgements, and theoretical considerations that we believe bear on accepting both the considered judgements and the principles'.

Critical scrutiny of the ethical 'data' in SA

As Plato's dialogues and Aristotle's *Nicomachean Ethics* reveal, inconsistencies on ethical matters abound among given views. This is also the background assumption in Daniels's 'working back and forth' idea, though not the main focus in Rawls's basically foundationalist IL. It is therefore most important to have some methodological ideas about how to work with inconsistencies to distinguish the true from the false. A great merit of SA is its heuristics for explicating and solving inconsistencies (SA 2). In this, Aristotle builds on the logical tools developed by his teacher and, like Plato's Socrates, applies the elenchus, although without an explicit question-and-answer structure, to argue conflicting views both for and against to detect their strengths and weaknesses.[67] Critical assessment also involves conceptual clarification and formulation of the given views in a clearer

[64] In his 'The Value of *Endoxa* in Ethical Argument' Klein focuses on the distinction between regulative and substantial *endoxa*, important in Aristotle's *eudaimonia* argument in *Nicomachean Ethics* I and in Rawls's *A Theory of Justice*. This distinction is not relevant in 'Two Concepts of Rules', so we shall not discuss it here.

[65] *Topics* I.1, 100a29–30; Barnes, 'Aristotle', pp. 498–50; R. Smith, *Aristotle: Topics Books I and VIII with Excerpts from Related Texts*, trans. with an Introduction and Commentary by R. Smith (Oxford, 1997), pp. 343–7; Marja-Liisa Kakkuri-Knuuttila, 'The Relevance of Dialectical Skills to Philosophical Inquiry in Aristotle', *RHIZAI: A Journal for Ancient Philosophy and Science* 2 (2005), pp. 31–74.

[66] For Aristotle *endoxa* are not only relevant in ethics, but in philosophy in general, as well as in dialectical and rhetorical argument.

[67] G. A. Scott, *Does Socrates Have a Method? Rethinking the Elenchus in Plato's Dialogues and Beyond* (University Park, Pa., 2002), pp. 19–35.

way to gain a better grasp of their implications. This is the first step in our attempt to clarify further Daniels's description of the research process leading to reflective equilibrium.

Generating the solution in SA

A typical feature of the Aristotelian SA is that the phases of critical scrutiny (SA 2) and constructing the solution (SA 3) are often closely related. Since the critical scrutiny aims to disclose what is true and what is false in the appearances, it simultaneously gathers information for the *construction of the solution* to the given problem. Sometimes the solution can be produced simply by combining what is left standing of the appearances after the critical phase, but this is obviously not always the case as innovative steps may be needed to construct an umbrella conception to 'save the appearances'. The simple case is engendered by two or more conflicting over-generalizations, since a successful argument for and against already reveals how to qualify each to create a synthesis.

Plato's famous method of collection and division in the *Phaedrus* is an early example of building a synthesis in the case of conflicting views.[68] The dialogue exemplifies this through Socrates' two speeches about love, the first one based on the hypothesis that love is a mad desire for the physical pleasure of beauty, and the other defining love as divine madness. The clue to the solution is the specification of an appropriate generic term (collection phase), here madness, and then dividing it into proper subclasses.[69] A somewhat similar technique is applied by Aristotle in his treatment of weakness of will in *Nicomachean Ethic*s VII.3, where the challenge is posed by Socrates' claim, against the majority view, that one cannot act against one's knowledge. Aristotle's solution, the details of which are much disputed by scholars, is based on several qualifications of the concept of knowledge.[70] To solve more demanding problems involves, in contrast, genuine moments of invention with no mechanical rules of discovery. However, the critical phase may offer significant help in determining where to seek the solution.[71] As we shall see, Rawls's treatment of

[68] *Topics*, 265a–266c.

[69] Plato, *Phaedrus*, trans. with an Introduction and Commentary by R. Hackforth (Cambridge, 1952); T. Calvo, 'Socrates' First Speech in the Phaedrus and Plato's Criticism of Rhetoric', *Understanding the 'Phaedrus': Proceedings of the II Symposium Platonicu*, ed. L. Rossetti (Sankt Augustin, 1992), pp. 47–60.

[70] R. Bolton, 'Aristotle on the Objectivity of Ethics', *Aristotle's Ethics*, ed. J. P. Anton and A. Preuls (Albany, 1991), pp. 7–28; David Charles, 'Nicomachean Ethics VII.3: Varieties of *akrasia*', *Aristotle: 'Nicomachean Ethics', Book VII Symposium Aristotelicum*, ed. Carlo Natali (Oxford, 2009), pp. 41–71.

[71] Aristotle's characterization of good human life (*eudaimonia*) in *Nicomachean Ethics* I.7 as activity according to the virtues could be mentioned as a solution which does not

the justification of punishment in 'Two Concepts' falls within the simple types of problem arising from conflicting over-generalizations. In developing one's capacity to solve theoretical problems of various kinds to generate reflective equilibriums, one may profit greatly by studying suitable literature, for instance Aristotle's treatises. This is the second move to specify Daniels's notion of research process leading to reflective equilibrium.

Justification in SA

The fourth and final phase in SA consists of the *justification of the solution* (SA 4) which, according to the excerpt quoted above, presupposes establishing that the problem is solved and that the appearances are saved. This also clarifies what, in fact, is involved in a reflective equilibrium of given conceptions, and its justification. Aristotle's requirement that the problems are solved clearly demands the basic condition of coherence, i.e. that the solution is consistent. The second justificatory task, responsible for the title *saving the appearances*, is to demonstrate that the initially conflicting appearances, or most of them, are included in the solution in a suitably revised form. In the case of qualified over-generalizations, this causes no particular difficulty, as already exposed with the help of the example from Plato's *Phaedrus*.

A further form of justification pointed out and sometimes applied by Aristotle, although not mentioned in the above passage, corresponds to Rawls's IL suggestion that, when considered judgements need a modification, the reason for the anomaly should be given.[72] Aristotle illuminates the importance of explaining why a false view arose in the first place as follows:

> We must, however, not only state the true view, but also explain the false view, since an explanation of that promotes confidence. For when we have an apparently reasonable explanation of why a false view appears true, that makes us more confident of the true view.[73]

'Saving the appearances' thus appears to lead to the kind of end result of research characterized by 'reflective equilibrium': the prior beliefs, or at least most of them when suitably modified, need to find their place in the end result. We hope that our presentation of the SA methodology has by now revealed that the four Aristotelian moves (SA 1)–(SA 4) offer a stronger heuristics for generating a reflective equilibrium than Daniels's three-step conception (D 1)–(D 3). For instance, in addition

follow directly from the critique of the current views, identifying good life with honour, wealth or pleasure.

[72] Rawls, 'Outline', p. 189.

[73] *Nicomachean Ethics* VII.14, 1154a22–25.

to generating hints for solving the problem in question (SA 3), the argumentative procedure in the critical scrutiny phase (SA 2) also yields material for justifying the solution in the manner according to (SA 4), and as required by coherence theory. We need to point out, however, that because of the research heuristics, SA is not to be identified merely as a form of coherence theory, since coherence theory in the standard sense is merely an epistemological theory of justification without elements of theory generation.

Rawls's IL compared with Aristotle's SA

The Aristotelian SA leading to a reflective equilibrium deviates fundamentally from the Rawlsian IL. The differences concern every phase of the research process, the understanding of what kind of 'data' to adopt as the starting point of ethical theorizing, the nature of problems arising from the 'data', the heuristics of generating ethical principles and the justification of the principles. The major cause of the differences is the logical structure of 'data', since in IL the 'data' are for the most part mutually consistent, whereas Aristotle focuses on conflicts within the 'data'. The Stagirite allows a much wider variety of 'data' as they may consist of various kinds of theoretical views held by almost everyone or by philosophers, while in IL the ethical 'data' comprises particular decisions made by competent judges concerning situations of conflict of interest.

It is worth noting that in spite of Rawls's strong scientistic ambitions, his considered judgements involve a strong Aristotelian element. Instead of being positivist-type observations, considered judgements made by competent judges are *endoxa* in the Aristotelian sense. Considered judgements are, first of all, not observations but judgements and, second, the competent judges are a particular subgroup of reputable people.[74] Aristotle would, however, relax the Rawlsian requirement that the competent judge needs to be a party external to the situation, since he accepts interest-laden decisions, which are to be saved by revealing their underlying interests.[75]

The intended illustrations in Rawls's dissertation and the 'Outline' essay include even stronger Aristotelian elements. The principles and general beliefs Rawls aims to justify and those he uses as evidence are without doubt Aristotelian *endoxa*. To explain why he seldom sees a need to begin by collecting considered judgements and explicating new principles, Rawls notes that 'the important theories of the past should be known, since it is highly possible that some one of them, or

[74] Klein ('The Value of Endoxa', p. 141) takes it for granted that the 'data' in Rawls's *A Theory of Justice* are *endoxa*.

[75] *Nicomachean Ethics* I.4–5.

some combination of them, may be correct after certain adjustments and changes have been made'.[76] Here he heavily but unconsciously undermines the reasonableness of his own strongly foundationalist IL methodology in favour of the reasonableness of the hypothetico-decuctive one. Interestingly enough, the remark bears close affinities to a comment by Aristotle:[77]

> We must consider it [characterization of *eudaimonia*], however, in the light not only of our conclusion and our premises, but also of what is commonly said about it; for with a true view all the facts harmonize, but with a false one they soon clash. . . . Now some of these views [concerning *eudaimonia*] have been held by many men and men of old, others by a few reputable persons (*endoxoi*); and it is not probable that either of these should be entirely mistaken, but rather that they should be right in at least some one respect or even in most respects.[78]

Aristotle is evidently expressing his belief in the reasonableness of the saving the appearances methodology.

To account for the differences between IL and SA, we could point out that Aristotle and the early Rawls emphasize contrary aspects in ethical research. While Aristotle highlights conflicts in given appearances and thus has a strong coherentist involvement, Rawls stresses consistency, implying a foundationalist emphasis. When working with more demanding theoretical problems, Rawls is also bound to face inconsistent appearances, as we shall see in the next section. Interestingly enough, his approach then has a strong Aristotelian feel.

METHODOLOGICAL ANALYSIS OF 'TWO CONCEPTS OF RULES'

Significance of 'Two Concepts of Rules'

We shall demonstrate in this section that Rawls's defence of utilitarianism in 'Two Concepts of Rules' when dealing with the justification of punishment conforms to the Aristotelian principle of saving the appearances. As has been shown in the preceding section the SA methodology offers a more detailed description of how to reach reflective equilibrium than Daniels's 'working back and forth' account. However, the essay merits a methodological analysis on its own for its importance among Rawls's works. Even though overshadowed by his main works, its significance is shown by the numerous translations as well as by its enduring position in anthologies dealing with moral

[76] Rawls, 'A Study', p. 69.
[77] For similar views, see Rawls, 'A Study', pp. 86, 109.
[78] *Nicomachean Ethics* I.8, 1098b9–29.

philosophy. It is incorporated into Rawls's *Collected Papers*, and has attracted interest in jurisprudence, sociology, political science, psychology, economics and organization studies. Quite recently the *Journal of Classical Sociology* published a special issue edited by Anne Warfield Rawls focusing on the theoretical relevance of the 'Two Concepts' to contemporary sociology.

As the title of the essay indicates, Rawls's main aim in his 'Two Concepts' is to argue for two notions of rules, namely, the summary and the practice conceptions. The distinction relates to two forms of ethical argument, 'justifying a practice and justifying a particular action falling under it'.[79] By applying the distinction to two cases, the justification of punishment and promise, Rawls aims to defend utilitarianism against some traditional objections by developing a version of it called 'practice-utilitarianism',[80] in which the rule-based conception of practice plays an important role, and helps to avoid certain typical objections relating to the justification of the institutions of punishment and promise. Rawls does not, however, aim at a complete defence of the utilitarian moral doctrine, but intends to make 'a logical point' which in itself 'leads to no particular social or political attitude'.[81] Thus the 'Two Concepts' essay helps us to understand Rawls's later attitude to utilitarianism.[82]

The 'Two Concepts' includes a revealing methodological ambiguity, however. Having briefly presented his solution to the justification of punishment (JP), Rawls remarks that the two theories of punishment, the utilitarian and the retributive view, have been reconciled 'by the time-honored device of making them apply to different situations'.[83]

[79] Rawls, 'Two Concepts' (1955), p. 3. Rawls notes himself that the distinction between justifying a practice and justifying a particular action has a long history, beginning with David Hume. He mentions, for instance, John D. Mabbot, 'Punishment', *Mind* 48 (1939), pp. 152–67 and James O. Urmson, 'The Interpretation of the Moral Philosophy of J. S. Mill', *Philosophical Quarterly* 3 (1953), pp. 33–9.

[80] In contemporary discussions 'rule-utilitarianism' seems to be a more familiar title than 'practice utilitarianism'. In rule-utilitarianism the rightness of particular acts depends on their conformity with the set of rules which, if generally accepted, would maximize the utilitarian conception of good. On different versions of utilitarianism see e.g. D. Lyons, *Forms and Limits of Utilitarianism* (Oxford, 1965); *Utilitarianism and Beyond*, ed. A. Sen and B. Williams (Cambridge, 1982); Derek Parfit, *Reasons and Persons* (Oxford, 1984); *Utilitarianism and its Critics*, ed. Jonathan Glover (New York, 1990); Geoffrey Scarre, *Utilitarianism* (London, 1996).

[81] Rawls, 'Two Concepts', pp. 4 and 32.

[82] Thus the essay sheds light on the guidelines Rawls later gave his students, according to which the precondition of a critique of a philosophical doctrine is the construction of the most reasonable interpretation of the doctrine in question: see Rawls, *Lectures on the History*, p. 18; A. Reath, B. Herman and C. M. Korsgaard, 'Introduction', *Reclaiming the History of Ethics: Essays for John Rawls*, ed. A. Reath, B. Herman and C. M. Korsgaard (Cambridge, 1997), pp. 1–5.

[83] Rawls, 'Two Concepts', p. 7. The importance of the justification of punishment is already mentioned by Rawls in the Rawls, 'Outline', p. 188.

In the same vein, he states that one of his aims is to formulate utilitarianism 'in a way which *saves* it from several traditional objections',[84] and 'allows for the sound points of its critics'.[85] These remarks reveal a connection with Plato's method of collection and division, further extended by Aristotle's SA. However, at the beginning of the essay Rawls refers to his IL methodology by pointing out that the aim of the paper is to 'state utilitarianism in a way which makes it a much better *explication* of our considered moral judgments than these traditional objections would seem to admit'.[86] For the meaning of 'explication', he refers to the 'Outline' essay.[87]

Because of the obvious discrepancy between the first two remarks and the third, our aim is to go deeper into the methodological practice of the essay by clarifying the structure of the JP argument in particular. We shall demonstrate point by point how Rawls's methodology in solving the conflict between the two theories of punishment in the 'Two Concepts' follows the lead of the Aristotelian SA. The notion of 'data', the nature of the research problem, and the way of discovery, as well as justification, accord with the SA rather than the IL methodology. This supports our claim that Rawls's 'Two Concepts' involves a move towards a new methodological practice, resembling his descriptions of RE. We shall next show how the four phases of SA offer a more satisfactory reading of the JP argument than an IL reading.

Collecting ethical 'data' in JP

At the beginning of his discussion of JP in the 'Two Concepts' essay, Rawls notes that, in spite of the shared agreement concerning punishment in the sense of attaching penalties to the violation of legal rules, the topic has been a troublesome one, the problem being the moral justification of punishment. He points out that none of the justifications has so far 'won any sort of general acceptance; no justification is without those who detest it'.[88]

Rawls next introduces two competing views on the justification of punishment, the retributive and the utilitarian views, in accordance with the first phase of SA. The retributive view holds that 'punishment is justified on the grounds that wrongdoing merits punishment' and the severity of the appropriate punishment should depend on the depravity of the wrongdoing. The utilitarian view, by contrast, looks forward, maintaining that justifiable punishment effectively promotes

[84] Rawls, 'Two Concepts', p. 32, our emphasis.
[85] Rawls, 'Two Concepts', p. 4.
[86] Rawls, 'Two Concepts', pp. 3–4.
[87] Rawls, 'Two Concepts', p. 4, n. 3.
[88] Rawls, 'Two Concepts', p. 4.

the interests of the society and that punishment is justifiable only by reference to its probable consequences.[89]

We may note immediately that the retributive and utilitarian justifications of punishment do not constitute the kind of knowledge basis for ethical decision-making as required by Rawls's IL. First of all, they are not decisions made spontaneously in moral cases of conflicting interests, but are, instead, general theoretical conceptions with their own justification. This implies that they satisfy the condition of being 'considered judgements' made by 'competent judges' in the Aristotelian sense of reputable opinions. Rawls himself states that both positions are worth serious consideration since 'intelligent and sensitive persons have been on both sides of the argument'.[90]

Critical scrutiny of the ethical 'data' in JP

In carrying out our analysis, it is helpful to note that the phase of critical scrutiny (SA 2) may already involve support for the solution, and no strict order of presentation is presupposed in SA. After a preliminary discussion of the two conceptions of punishment, Rawls proceeds to the solution,[91] and then, in order to offer further justification to the solution suggested, continues by exploring some objections to utilitarianism presented from the retributionist point of view.

To start with, Rawls notes that there is a contradiction between the retributive and utilitarian views by pointing out simply that the way each is stated makes 'one feel the conflict between them'.[92] Having next suggested his solution to the conflict, he focuses on the retributionists' challenge of 'whether utilitarianism doesn't justify too much'.[93] Isn't the utilitarian committed to accepting the punishment of innocent people, for instance, in a situation where society is shocked by a terrorist attack? According to the retributionist, it is justifiable to punish only the real terrorists, while he could argue that a utilitarian might punish justifiably innocent people if that would effectively promote the interests of the society. This could be the case, for instance, when the real terrorists are extremely difficult to catch, and the interests of the society are advanced when some person or a group is charged and punished credibly and publicly for the attack. Isn't it in the interest of society to instil fear in future terrorists and, more importantly, to restore the feeling of security among its people?

[89] Rawls, 'Two Concepts', pp. 4–5.
[90] Rawls, 'Two Concepts', p. 6.
[91] Rawls, 'Two Concepts', pp. 4–8.
[92] Rawls, 'Two Concepts', p. 5. Here Rawls also claims to have laid out the appearances in a way that the reader starts to wonder 'how they can be reconciled'.
[93] Rawls, 'Two Concepts', p. 8.

In his defence, Rawls argues that such considerations are excluded for the utilitarian on the basis of the concept of punishment: 'utilitarians agree that punishment is to be inflicted only for the violation of law'.[94] This is thus an assumption shared by both parties. For closer consideration of the retributionist stand, he cites a lengthy passage from Carritt, which begins: 'the utilitarian must hold that *we* are justified in inflicting pain always and only to prevent worse pain or bring about greater happiness'.[95]

We need not cite the passage in full, since this suffices to expose Rawls's objective. He rejects Carritt's criticism as futile because of the failure to specify the pertinent institution or institutions referred to by the ambiguous 'we'.[96] This observation prepares the ground for constructing the solution, as sometimes happens at the critical scrutiny phase (SA 2).

Generation of the solution in JP

In accordance with the third phase of SA, Rawls shows that limiting the scope of application of the utilitarian and retributive views on justification reconciles them so that they no longer contradict each other. Following the suggestion to specify the agents and institutions relevant to punishment, he draws the distinction between 'justifying a practice as a system of rules' and 'justifying a particular action which falls under these rules'.[97] This yields the so-called practice-utilitarian view, the proper scope of which is the justification of the social practice of punishment as a system of rules. The retributive view, in contrast, is appropriate with regard to questions about 'application of particular rules to particular cases'.[98]

In a democratic society there are, in fact, two distinct institutional actors associated with punishment, the legislator and the judge. The forward-looking utilitarian view is, hence, the viewpoint of the legislator, whereas the backward-looking retributive view is the viewpoint of the judge.[99] When performing the office of a judge, a person cannot adopt some sort of utilitarian legislator's viewpoint and distribute punishment as a means of serving the interests of society. A judge should strive to achieve a state of affairs in which only the guilty are punished. This way of resolving the conflict by means of

[94] Rawls,' Two Concepts', p. 7.
[95] Rawls, 'Two Concepts', p. 10, our emphasis.
[96] Rawls, 'Two Concepts', pp. 10–13.
[97] Rawls, 'Two Concepts', pp. 5, 32.
[98] Rawls, 'Two Concepts', p. 5.
[99] Rawls, 'Two Concepts', pp. 6–7. We find the distinction between the backward-looking forensic argument and the forward-looking political argument in Aristotle's *Rhetoric* I.3.

specifying a general concept and then dividing it appropriately closely resembles Plato's methodology of collection and division, and thus explains Rawls's statement that he 'reconciles the two views by the time-honored device of making them apply to different situations'.

Justification of the solution in JP

Finally we shall show how, in his justification of the solution, Rawls adopts the various argument forms typical of SA. Rawls rejects the retributionist's worry that a utilitarianist might accept that punishing the innocent exemplifies justification for utilitarianism drawn at the critical phase (SA 2). Showing that the conflict in the appearances has been resolved is achieved by confining each position to a particular social context, the retributive conception to the judge and the utilitarian to the legislator. This is a way of saving the appearances, as noted by Rawls himself when stating that the solution simultaneously 'allows for the apparent intent of each side' and 'seems to account for what both sides have wanted to say'.[100] The retributionists have correctly insisted 'That no man can be punished unless he is guilty, that is, unless he has broken the law'.[101] The utilitarians, by contrast, are concerned with the institution of punishment as a system of rules to foster the good of society effectively.[102]

Rawls further illustrates the soundness of the solution with the example of a boy putting two distinct explanatory demands to his father. In one, the boy requires the explanation of the imprisonment of a particular person, and in the other the explanation of the institution of punishment. To clarify the difference between these two questions, Rawls formulates them with the help of illuminating contrasts.[103] The first question asks 'why was J punished rather than someone else?' and the second asks 'why do people punish one another rather than, say, always forgive one another?'[104]

Rawls furthermore applies the argument form included in both IL and SA by suggesting how one can miss the distinction between justifying a practice and justifying a particular action falling under it. This is where the two concepts of rules, the practice and the summary conception, step in. According to the former, rules constitute a practice by defining what actions are appropriate to the practice in question while, according to the latter, rules are pieces of advice to speed

[100] Rawls, 'Two Concepts', pp. 7–8.
[101] Rawls, 'Two Concepts', p. 7.
[102] Rawls, 'Two Concepts', p. 8.
[103] For the recent notion of the *contrastive concept of explanation*, see Petri Ylikoski, *Understanding Interests and Causal Explanation*, <www.ethesis.helsinki.fi/julkaisut/val/ kayta/vk/ylikoski/understa.pdf/> (PhD Dissertation, Helsinki, 2001).
[104] Rawls, 'Two Concepts', pp. 5–6.

up decision-making in particular cases that tend to recur. A major difference between these two kinds of rule is the order of logical priority. While practice-rules are logically prior to particular actions, particular actions are logically prior to the summary rules:

> in a practice there are rules setting up offices, specifying certain forms of action appropriate to various offices, establishing penalties for the breach of rules, and so on. . . . given any rule which specifies a form of action (a move), a particular action which would be taken as falling under this rule given that there is the practice would not be *described as* that sort of action unless there was the practice.[105]

Thus it is logically impossible to perform a particular action outside the particular practices setting the stage for it. For instance, getting a free-kick in a game of soccer is possible only through playing the game in question, i.e. by following the rules which define the game.[106]

Even though summary rules, or what could also be called 'strategic rules', may contain historically tested practical knowledge, they are only guides for decision-making, and there is nothing wrong in questioning their applicability to particular cases. For instance, it is possible to play soccer by adopting strategic rules (strategies) that have proved successful in previous games. These strategic rules do not, however, define soccer as a game, because it is possible to play the game without adopting them.[107] Rawls emphasizes, however, that not all rules fit nicely into these two categories, or that only one of these conceptions is the right view.[108] According to him, it is the summary conception of rules that fails to perceive the significance of distinguishing justifying a practice as a system of rules and justifying particular actions falling under it, while the practice conception stresses the significance of the distinction.[109]

Summary of the analysis of JP

We have shown that Rawls's JP argument in the 'Two Concepts' accords neatly with the four stages of Aristotle's SA methodology rather than with the three phases of his IL methodology developed in the dissertation and repeated again in the 'Outline' essay. Clearly, the starting point of the investigation, the retributive and the utilitarian forms of justifying punishment, are not decisions made by competent judges in situations of moral conflict of interest, as the 'data' characterized in IL (IL 1) are theoretical principles worth serious

[105] Rawls, 'Two Concepts', p. 25.
[106] Rawls, 'Two Concepts', p. 25.
[107] Rawls, 'Two Concepts', pp. 22–4.
[108] Rawls, 'Two Concepts', p. 29.
[109] Rawls, 'Two Concepts', pp. 19, 22–4.

consideration, corresponding to Aristotle's *endoxa* (SA 1). Since these views are in conflict, the task for the philosopher cannot be simply to explicate a more general conception to cover them (IL 2). SA has an adequate approach to such problem situations by suggesting that one examine both views critically before attempting a solution to the conflict. To extract the truth involved, Rawls proceeds by articulating the strengths and weaknesses of both positions, as if following the Aristotelian guidelines (SA 2). Having constructed a synthesis of them (SA 3), he justifies this by demonstrating that, in addition to having resolved the conflict, the intention of the proponents of both views is saved. Furthermore, he explains the origin of the conflict (SA 4).

Perhaps the most striking divergence between the Aristotelian SA and the Rawlsian IL methodology concerns the heuristics of discovery. While Rawls explicitly speaks about 'explication' as a 'heuristic device', his characterization of 'explication' says very little about how the principles are to be constructed. Resolving conflicts between prevailing theoretical conceptions through their criticism offers a fairly strong heuristics for creating theoretical ethical principles at any rate, if not a 'logic' or 'mechanical methodology' of discovery.[110] This concludes our argument that Rawls's actual methodological practice in 'Two Concepts' follows the Aristotelian methodology of saving the appearances (SA) rather than his inductive logic methodology (IL) of the dissertation. Our analysis of the argument structure of his manner of solving the problem of justification of punishment illuminates how the SA methodology yields a richer description of discovery and justification than Daniels's (D 1)–(D 3).

CONCLUSION

The chief purpose of this article has been to investigate Rawls's early methodological thinking and practice. We began by presenting what we call his *inductive logic* (IL) methodology, laid out in his dissertation 'A Study in the Grounds of Ethical Knowledge: Considered with Reference to Judgments on the Moral of Worth of Character', and repeated in 'Outlines of a Decision Procedure for Ethics'. Rawls created his IL in simulation of the newly developed inductive logic with the ambition to find a methodology, convincing to positivistically minded thinkers as well, for expounding principles to offer help for ethical decision-making.

Our analysis has revealed that Rawls's IL is primarily based on a foundationalist epistemology with a coherence-theoretical element.

[110] Rawls, 'A Study', pp. 68–9; Rawls, 'Outline', pp. 178, 184.

However, its intended illustrations fail to demonstrate that the three-phase research process prescribed by IL is a reasonable methodology for producing and justifying ethical principles. The examples both in the dissertation and in the 'Outline' essay lack the phase of collecting the considered judgements made by competent judges in situations of conflict of interest (L 1), and the phase of explicating the general principles on the basis of the considered judgements (L 2). Instead of an inductivist methodology, the examples illustrate a hypothetico-deductive approach by testing a given set of principles, and thus merely involve the last stage of justification (IL 3). More importantly, the evidence supplied does not satisfy the criteria for appropriate 'data' (IL 1); in fact, the 'data' consists of abstract theoretical principles rather than actual decisions made by competent judges in situations of conflict of interest as required by IL. To show that Rawls's IL may still be valuable for generating ethical principles, we presented one example from Aristotle's *Nicomachean Ethics*.

The examination of the justification of punishment argument (JP) in the somewhat later 'Two Concepts of Rules' essay reveals an even greater move away from the IL methodology, implying a significant turning point in Rawls's methodological practice. The positive part of our investigation is the identification of Rawls's argument structure in his defence of utilitarianism in the 'Two Concepts' with the Aristotelian methodology of saving the appearances (SA). With the help of the SA methodology we have shown that Rawls ends up in a reflective equilibrium of the utilitarian and retributionist justifications of punishment by saving the truth of both conceptions by suitably qualifying them. This solution corresponds neatly with his later characterizations of RE in the following terms:[111]

> Justification is a matter of mutual support of many considerations.[112]

> [T]here are no judgments on any level of generality that are in principle immune to revision.[113]

The particular value of the Aristotelian SA lies in the fact that, in comparison to Daniels's account of the Rawlsian RE methodology, SA yields a deeper understanding of RE by offering a more detailed

[111] See also Rawls, *A Theory*, pp. 48–9; Rawls, 'The Independence', pp. 28–9; Rawls, *Lectures on the History*, p. xvi; Rawls, *A Restatement*, pp. 29–32. The considered judgements are, according to Rawls, in reflective equilibrium when they have been brought into harmony with the new theoretical principles: see Rawls, *A Theory*, pp. 46, 50–1. Here again the 'considered judgements' need to be understood as Aristotelian *endoxa*.

[112] Rawls, *A Theory*, pp. 21, 579.

[113] Rawls, 'The Independence', p. 289.

description of how to reach and justify a reflective equilibrium. This implies that SA is not a purely coherence-theoretical view of justification, as it includes a heuristics of discovery as well. Thus we venture to suggest that the Aristotelian saving the appearances is a handy methodology to reach the state of reflective equilibrium in cases of individual theoretical problems.[114]

We may point out further that the change in methodological practice in the 'Two Concepts' essay corresponds to the substantial innovation of distinguishing two concepts of social rules, i.e. summary rules and rules of practice. Rawls's IL approach to ethical theorizing and his hypothetico-deductive examples in both the dissertation and 'Outline' are tied to the summary conception of ethical rules. This kind of inductivist or more generally statistical approach to social reality is inappropriate when exploring constitutive social rules, which requires holistic and contextualist research methodologies.[115] The 'Two Concepts' essay thus marks a turning point not only in Rawls's methodological practice but in his theoretical thinking more generally. The explicit ambiguity in Rawls's methodological comments in 'Two Concepts' indicates, however, that at the time he was not yet quite aware of the new methodological direction, and perhaps also not quite aware of the wider theoretical goals indicated in the essay.

Our investigations have revealed that as regards the timing of Rawls's adoption of some version of RE, the date has to be moved from *A Theory of Justice* to some fifteen years earlier to the 'Two Concepts' essay. We have shown further that RE is not a recent invention, and its roots go all the way back to ancient philosophy and Aristotle. As our analysis of the justification of punishment argument in the 'Two Concepts' indicates, the SA may serve exactly the kind of practical aims of political philosophy imposed by Rawls's later works. Applying the SA methodology in a satisfactory manner one may resolve crucial divisive political conflicts by producing a *reflective equilibrium* between the parties to the conflict.[116] To what extent SA corresponds to Rawls's methodological practice in his mature works and to what extent it

[114] Nussbaum, 'Equilibrium'; Daniels, 'Reflective Equilibrium'; Daniels, *Justice and Justification*, pp. 1–2.

[115] Anne Warfield Rawls, 'An Essay on Two Conceptions of Social Order', *Journal of Classical Sociology* 9 (2009), pp. 500–20.

[116] J. Rawls, *Justice as Fairness: A Restatement*, ed. E. Kelly (Cambridge, MA, 2001), pp. 1–2.

helps to elucidate the disputes concerning RE remains to be assessed in future research.[117]

jukka.makinen@aalto.fi
marja-liisa.kakkuri-knuuttila@aalto.fi

[117] The authors are indebted to Brad Hooker and an anonymous reviewer for their insightful and very helpful reviews and guidance. We would like to express our gratitude to Jeroen van den Hoven, Hasse Hämäläinen, Rex Martin, Martha Craven Nussbaum, Kristina Rolin, Walther Schweidler, Ville Päivänsalo, David Rönnegard and Juha Sihvola for support and valuable comments on earlier versions of the article. We gratefully acknowledge the financial support of the Academy of Finland and Liikesivistysrahasto (Foundation for Economic Education).