Three Remarks on "Reflective Equilibrium":
On the Use and Misuse of Rawls' Balancing Concept in Contemporary Ethics

Dietmar Hübner

Abstract

John Rawls' notion of "reflective equilibrium" is one of the most utilised concepts in contemporary ethics when it comes to the basic methodological question of how to establish and trade off different normative positions and attitudes. Even where Rawls' specific contractualist account is not adhered to, "reflective equilibrium" is readily adopted as the guiding idea behind coherentist approaches, which seek moral justification not in a purely deductive or inductive manner, but in some balancing procedure that will eventually produce a stable adjustment of relevant doctrines and standpoints. However, it appears that the widespread use of this idea has led to some considerable deviations from its meaning within Rawls' original framework and to a critical loss of conceptual cogency as an ethico-hermeneutical tool. This contribution identifies three kinds of "balancing" constellations that are frequently but erroneously brought forward under the heading of Rawlsian "reflective equilibrium": (a) balancing theoretical accounts against intuitive convictions; (b) balancing general principles against particular judgements; (c) balancing opposite ethical conceptions or divergent moral statements, respectively. It is argued that each of these applications departs from Rawls' original construction of "reflective equilibrium" and also deprives the idea of its reliability in clarifying and weighing moral stances.

Keywords: John Rawls, reflective equilibrium, contractualism, coherentism

I. Introduction

(1) When ethicists are confronted with the fundamental question of how to decide between different notions of moral behaviour or just institutions, John Rawls' "reflective equilibrium" is frequently evoked as a suitable instrument for balancing relevant positions and attitudes and reaching a stable and informed trade-off between them. Even where Rawls' explicitly contractualist design is not adopted, "reflective equilibrium" is readily assumed, constituting the common denominator of so-called "coherentist" approaches to ethics. Such positions do not restrict themselves to deductive arguments (starting off from established axioms and deriving their ethical implications) or inductive accounts (picking up punctual insights and inferring their ethical backdrops) but try to combine both pathways in order to achieve some kind of mutual adjustment between different levels of moral reasoning. Applied ethics, in particular, with its constitutive horizon of new ethical challenges and an inhomogeneous supply of alternative ethical principles, has made intensive use of the idea of "reflective equilibrium". This should purportedly be established between various moral aspects of a given situation (e.g. expressed wishes, medical prospects, limited

resources) or the ethical tenets that endorse them (such as autonomy, beneficence, or justice). Consequently, "reflective equilibrium" has become a key term in fundamental reflections on biomedical and technological ethics, a guiding concept for ethics committees and review boards and, not least, a standard reference in methodology sections of funding proposals in these growing areas.

However, it appears that this frequent appeal to "reflective equilibrium" is, at times, unjustified when compared to the primary idea that Rawls brought forward under that title and the specific function it fulfilled within his philosophical framework: at least in some realms of contemporary ethics the use of "reflective equilibrium" seems to have become inflationary. This is not meant in the harmless sense of "widespread", but in the full sense of "devaluing": "reflective equilibrium" is quoted in connections that are not warranted by its original meaning, and it is applied in ways that infringe upon its erstwhile conclusiveness.

The point of my criticism is not merely exegetical: it does not confine itself to demonstrating that prevalent references to Rawls illegitimately attempt to justify their lines of argument by appealing to an authority from which they in fact considerably deviate. My primary aim is rather categorical in nature: I want to argue that these deviations transfer Rawls' concept to problems and constellations where its application lacks a sufficiently sound foundation.

(2) In the following sections I will, first, briefly sketch the methodological location of "reflective equilibrium" (henceforward: RE) in Rawls' original approach (II). This outline will help to preclude some basic fallacies concerning the function and import of the concept. Three subsequent sections will be devoted to specific interpretations of RE that seem to be predominant in contemporary ethics: RE as a concept for balancing *theoretical accounts* against *intuitive convictions* (III), for balancing *general principles* against *particular judgements* (IV), and for balancing *opposite ethical conceptions* or *divergent moral statements*, respectively (V). In each case I will demonstrate that the suggested interpretation of RE substantially departs from the way in which Rawls introduced the concept and that this variance deprives the idea of its reliability in clarifying and weighing normative statements. I conclude (VI) that RE is an ethical implement which may be successfully applied in a well-defined conceptual framework, connecting different levels of normative reasoning in a clear-cut manner and thus allowing for their hermeneutical deepening and reciprocal balancing. However, the establishment of such a framework requires accepting certain basic normative commitments, such as the ethical relevance of a contractualist thought experiment. Correspondingly, RE is never a "neutral" device, appropriate for balancing just any rivalling normative convictions, but rather a "relative" instrument, depending on the explicit approval of a particular ethical scaffold in which it may unfold its unequivocal effect.

It is foreseeable that the above program will irritate some readers—not least those who are well-acquainted with Rawls' theory and the development of his thought. For it seems obvious that the account that Rawls himself and a larger part of his most renowned commentators give of RE embraces exactly some of the "mistakes" that I claim to identify—most evidently the balancing of general principles against particular judgements, but also, to a lesser degree, the adjustment of theoretical accounts and intuitive convictions. Anticipating reactions of this kind, I would like to make the following disclaimer: I am *fully aware* that

some of Rawls' own works, prior to and after "A Theory of Justice", as well as many writings by other contributors, including his best-known followers, *do in fact* suggest uses of RE along the lines I want to criticise. In his earlier work, Rawls sketches a method of ethical decision-making that relates general principles of a non-intuitive kind to intuitive appraisals of particular cases [Rawls 1951, 181–190]. When establishing RE in "A Theory of Justice" he is inspired by Nelson Goodman's epistemological work on the mutual adjustment of general rules and particular inferences in deductive and inductive reasoning [Goodman 1954, 65–68]. In later explications of RE and, specifically, in suggesting the move from "narrow" to "wide" RE, Rawls explicitly allows for some of the interpretations I reject, particularly for the balancing of moral judgements at all levels of generality [Rawls 1974/75, 7–10; Rawls 2001, 29–32]. This idea of "wide" RE, embracing a most extensive range of moral and even non-moral beliefs and attitudes to be adjusted against each other, was elaborated most notably by Norman Daniels, whose work has become a standard reference for interpreting RE in contemporary ethics [Daniels 1996, 1–10, 22–46].

Notwithstanding these connections and developments, it seems worth noting that the way *John Rawls* eventually *defines* and, above all, *uses* the idea of RE in his *groundwork "A Theory of Justice"* considerably deviates from these broader understandings that are implied by the prior and later work of himself and others: RE in "A Theory of Justice" manifestly refrains from those wider applications and restricts itself to a much more confined and succinct meaning. This historical observation is interesting in its own right, as "A Theory of Justice" is, after all, the locus classicus and Rawls' major application of RE. But it may have systematic implications, too, suggesting that this narrower conception of RE is the more well-grounded and reliable one.

(3) Presumably, the extension of RE to a broader meaning and, particularly, the shift from "narrow" to "wide" RE was countenanced by complaints that the original idea was too reliant on a specific ethical framework (Rawls' contractualist approach) and, correspondingly, too restricted in its balancing components (choice parameters and principles derived). It was tempting, therefore, to extend RE beyond its prior limits and content in order to counter these accusations. Advocating "wide" RE promised to demonstrate that one had finally come to possess an argumentative device which was completely independent of any foregone ethical presumptions and suitable for the assessment of any moral stances whatsoever, thus minimising the risk of reproducing and perpetuating existent prejudices in one's primary outlook and starting convictions. I want to suggest, however, that it is exactly the embedment and restraint of RE in its original presentation that constitutes its chief virtue. Depriving RE of its definite foundation and overrating the array of stakes that it is able to balance will eventually vitiate its quality as a well-defined ethical tool for non-arbitrary progressive reasoning. It is the modesty of its account and application, acknowledging its systematic place within a specific line of argument and restricting it to the adjustment of the corresponding parts, that bestows reliability upon RE, as I shall try to show (for related criticisms of "wide" RE see e.g. Holmgren 1989, 58–59; Scanlon 2003, 150–151).

The main purpose of this paper is thus to defend Rawls' design and application of RE in "A Theory of Justice" against its various extensions and extrapolations in contemporary ethics. In doing so, I readily accept the charge of being "more Rawlsian" than the

Rawlsians, or even than Rawls himself, by sticking to a version of RE that he and his followers actually tried to transcend. I find this an unproblematic position as long as it serves to underline the conceptual clout that the idea displays in its more confined formulation. I am convinced that RE is a remarkable and reliable instrument for ethical reasoning. This paper, in large part, is meant to highlight its fruitfulness and conclusiveness. However, exactly this quality shows up most clearly when it is demonstrated *how well* RE works in its original narrow setting and *how much* of its conceptual vigour is lost once it is transferred to overly wide applications.

II. "Reflective equilibrium" in Rawls' "A Theory of Justice"

(1) It seems advisable to briefly recollect the precise function of RE in Rawls' original account (for an accurate summary see Tersman 1993, 124–135). In his groundwork "A Theory of Justice" (1971/99)[1] Rawls develops an explicitly contractualist approach to political theory, introducing a fictitious choice situation (the "initial situation") whose virtual participants are invited to freely select the principles of justice, defining the basic structure of the society in which they are going to live from that point on [ToJ, § 3, 10–11, § 10, 47, § 20, 102]. Their *cognitive and motivational features* are supposed to amount to mere rational self-interest, enabling them to anticipate the effects of different social arrangements on their future living conditions while restricting their intentions to the desire of maximising their own respective benefits [ToJ, § 3, 12, § 22, 111–112, § 25, 123–125]. At the same time, their *state of information* is limited (by a so-called "veil of ignorance") in such a way that they do not know their own future position, including their natural endowments and interests as well as their social status and circumstances, and thus cannot make their choices in a manner that would straightforwardly optimise their individual outcomes [ToJ, § 3, 11, § 3, 14, § 24, 118].

It is the core conviction of Rawlsian contractualism that this combination of self-interested rationality and limited information, guiding a hypothetical free choice of basic social principles, is suitable to deliver an ethical standard for real societies (in his terms: a result implementing "justice as fairness" [ToJ, § 3, 10–12, § 25, 128–129]). Of course, there is deep disagreement concerning the validity of this idea. Libertarians like Robert Nozick complain that reasoning from rational choices of self-interested agents is biased towards substantial modes of central distribution, as these agents will only be concerned with the probable outcomes of alternative social arrangements, while ignoring the inherent procedural justice of free market transactions [Nozick 1974, 198–204]. Communitarians like Michael Sandel deny that arguments referring to the fictitious choices of abstract voters, devoid of natural characteristics and social background, detached from every notion concerning a good life and any obligations based on human relationships, are epistemologically sound and morally meaningful [Sandel 1982, 15–28, 59–65, 133–135, 147–154, 173–174, 175–183]. Egalitarians, utilitarians and Marxists, as well as discourse ethicists, democracy theorists and struc-

[1] In the following, I refer to the 1999 edition of "A Theory of Justice" [Rawls 1971/99] with the shortcut "ToJ", paragraph and page number.

turalist thinkers, have their specific objections against Rawls' model, preferring alternative forms of political reasoning. I will not engage in this fundamental discussion on the cogency of contractualist thinking in general, or Rawls' version of contractualism in particular.[2] But already at this stage of argument I want to endorse the idea that at least *some kind of conceptual framework*, like Rawls' contractualism, seems to be indispensible in order to make RE a reliable instrument in the first place: if there is to be any balancing procedure that provides not just an arbitrary comparing and selecting of alternative suggestions, but a guided deepening and adjustment of correlating arguments, it requires *a definite setting and overarching arrangement* of the components that are to be weighed against each other. This framework may be contractualist in nature or not. But in any case, if RE is supposed to be an efficient tool in ethical reasoning, it needs to be applied within some well-defined argumentative structure that establishes a solid and expandable connection between the different levels that it is meant to adjust.

(2) In Rawls' approach, this connection is established between the essential features of the *choice situation* designed and the *principles of justice* that rational, self-interested participants would select under these circumstances. Correspondingly, its argumentative substance is based on the fundamental outlook of *contractualist reasoning* and on the specific tenets of *rational choice theory*. Certain types of choice problems suggest associated decision rules, such as EU-maximisation (i.e. choosing the option with the best expectation value on a utility scale) or the maximin rule (i.e. focussing on the least favourable outcome in each option and choosing the option with the best of these worst outcomes). Within the given thought experiment, these rational decision rules, more or less straightforwardly, translate into principles of justice, e.g. into utilitarianism (more precisely into maximizing average utility rather than the sum total of utilities) or into the "difference principle", respectively (essentially optimising the position of the worst-off as the sole justification for social and economic inequalities). It is these two layers of Rawls' contractualist account that *are connected* by his framework and that are meant *to be balanced* against each other. RE, consequently, is supposed to mark precisely *the type of choice situation that is to be assumed* in the contractarian thought experiment and *the principles of justice that would ensue from it* through the standards of rational decision-making.

There has been considerable discussion with regard to the choice principles suggested by Rawls and other contractualists employing this kind of reasoning. EU-maximisation has been challenged by authors who regard classical utility theory as generally defective and try to develop alternative rules for rational decision-making [Allais 1953, 503–505, 527; Rescher 1983, 44–83, 114–119; Shackle 1955, 36–55, 63–67]. The maximin rule, in turn, has been criticised as not rationally unique, as overly conservative, as excessively pessimistic, etc., even with respect to the peculiar choice situation that Rawls envisages [Barber 1975, 297; Dworkin 2002, 330–331; Hare 1973, 106]. Again, I will not elaborate on these intricate decision-theoretic questions in what follows.[3] For the purposes of this contribu-

[2] I have argued elsewhere that modern contractualism of the Rawlsian type establishes one out of three essential approaches to distributive justice, each of them being well-founded in its theoretical setup but none of them having exclusive access to the problem [Hübner 2009].

[3] I have discussed the decision-theoretic details at length, concluding that EU-maximisation is actually deficient for decisions under risk, whereas the maximin rule may indeed pass for a rational choice criterion in decisions under uncertainty [Hübner 2001].

tion it suffices to remark that the proposed framework, in any case, makes up a *suitable basis* for seeking RE between its components: Rawls' approach does establish a clear-cut *deductive/inductive relationship* between the two levels of his argument, i.e. the characterisation of the choice problem and the selection of the principles of justice. The latter are deduced from the former, the former is induced by the latter, by way of their formal relationship within contractualism and their material connectivity provided for by decision theory. This relationship is not compromised by the fact that the uniqueness of certain decision rules in specific choice situations may be disputed. Such dispute concerns the concrete results of the construction. It does not challenge its basic functionality.

(3) However, this deductive/inductive relationship still contains an inherent plurality of possible solutions that is not reducible by any intensified discussion of adequate decision rules. The reason for this remaining diversity is that certain parameters of the "initial situation" are open to variation. Notably the "veil of ignorance", precluding information from the fictitious participants, may be configured in different ways. Of course, *knowledge of their individual positions* in the future society must be precluded in any case. Otherwise each participant would opt for those principles that best furthered his particular prospects. If that should happen the procedure would yield no solution, as the principles are supposed to be selected unanimously by the participants (a counterfactual majority vote would neither have a systematic ethical foundation within the model, nor would it permit a univocal philosophical reconstruction of its result). But this restriction still allows for *different amounts of knowledge* that the participants are given. They do not know their own respective positions, but they may have more or less insight into the possible societies that would be put into effect by the principles of justice between which they have to choose. They may know specific details about these societies, or they may be denied this particular knowledge (the "thickness" or the "veil of ignorance" may still differ).

The most prominent alternatives are the following. First, the participants may not know their *own positions* in the future society, but they will at least know the *relative frequencies* of different roles that this society has to offer. In this case, they find themselves in a situation of *decision under "risk"*, being able to calculate the probabilities of their various possible fates (assuming that there is some kind of lottery that assigns them their positions in society when the "veil of ignorance" is lifted, each single position having the same probability). Second, they may not even be given *these frequencies*, but only the *distinct levels* of well-being or prosperity that occur in the society in question. In that case, they are faced with a situation of *decision under "uncertainty"* in which they only know the possible outcomes that may await them, but not the associated probabilities (which is actually the minimum level of knowledge required for any rational decision procedure).

Prior to Rawls, John Harsanyi outlined a very similar contractualist argument and, opting for the "risk" version of the "initial situation" while claiming that EU-maximisation was the adequate rule for this kind of choice problem, derived average utilitarianism along the lines sketched above [Harsanyi 1953, 434–435; Harsanyi 1982, 44–48]. Rawls, by contrast, prefers the "uncertainty" version and, turning to the maximin rule for this type of choice situation, arrives at his "difference principle" [ToJ, § 26, 130–134, § 28, 149]. As I already

mentioned, both derivations have been criticised for decision-theoretic reasons.[4] But even without this disagreement, the model would still be ambivalent as long as there is no reproducible way to decide which version of the "initial situation", "risk" or "uncertainty", should actually apply. It is *exactly here* that the idea of RE comes in: RE, as introduced by Rawls, is supposed to determine which version of the "initial situation" and, correspondingly, which principles of justice should be chosen. It is based on the deductive/inductive relationship that holds between these two levels, the primary modelling of the "initial situation" and the ensuing principles of justice, and it is strived for in order to indicate which of their interrelated alternatives is preferable [ToJ, § 3, 14]. Each of them can be fixed in order to define the other. The model may be chosen to derive the ensuing principles (deductive line), or the principles may be chosen to select the fitting model (inductive line). RE is supposed to indicate a *mutual adjustment* of both layers, yielding their most convincing, stable and justified correspondence—in Rawls' terms: the "original position" (the appropriate "initial situation" [ToJ, § 3, 11, § 4, 15]) and the set of principles and priority rules that he finally endorses (as most adequate solutions to the underlying idea of free choice in a fair situation, i.e. to "justice as fairness" [ToJ, § 3, 11, § 4, 15]) [ToJ, § 4, 18, § 20, 105, § 87, 507].

This rough outline of the systematic place and argumentative function of RE in Rawls' "A Theory of Justice" will help in discussing and criticising alternative uses of the concept in contemporary ethics. One preliminary remark, however, is necessary at this stage as it forestalls some very superficial readings of RE that can occasionally be observed in literature: RE is not a *result* of Rawls' theory but an *instrument to develop* his theory. It is an ethico-hermeneutical device to generate a most convincing constellation of underlying model and principles implied, and it is only this final constellation of the two levels that Rawls eventually proposes as his theory. So RE is not Rawls' *principle*, but his *method to find* his principles. These principles are the "difference principle"—or more precisely: the full set of principles and priority rules [ToJ, § 11, 52–56, § 46, 266–267] which, he supposes, follow straightforwardly from his uncertainty interpretation of the "original position" and the maximin rule that prevails therein [ToJ, § 26, 132, § 29, 153].

It seems justified to recall this status of RE in Rawls' original approach when noting that contemporary ethicists at times finish their work in *calling for* RE to be established within the plurality of normative stances they investigate. On Rawls' account this kind of conclusion implies that the ethicist simply did not do his job, for it is *precisely he* who should make use of the concept in order to finalise his theory. The failure to do so is excusable in *meta-ethical* accounts e.g. of applied ethics. Here, RE may be quoted and endorsed as the proper tool for normative appraisals, its actual use being legitimately left to the *applied ethicist* in his work on concrete moral problems. More frequently, however, the idea seems to be that the application of RE was not the business of the *ethicist* at all: rather, the *actors* are supposed to establish RE between their various options, the ethicist providing nothing but abstract explanations concerning the meaning and function of that concept. This attitude is deeply at variance with the role that RE attains in Rawls. There, RE is the instrument that the ethicist

[4] Rawls basically agrees that the "risk" version would suggest EU-maximisation [ToJ, § 20, 105, § 27, 139–144], while Harsanyi remains negative against the maximin rule even for "uncertainty" situations [Harsanyi 1975, 37–40].

is obliged to use when elaborating his theory (and that Rawls himself does use to work out "A Theory of Justice"). Moreover, it is not uncommon that this deficient appreciation of RE's role in fundamental ethical reasoning (as opposed to concrete moral decision-making) goes along with exactly those misrepresentations that the remainder of this paper is devoted to: applying RE between theories and intuitions, between general principles and particular judgements, or between divergent moral statements on the same systematic level.

III. Balancing theory against intuition?

(1) RE is frequently conceived of as the proper adjustment of *theoretical accounts* (based on scholarly argument, systematic elaboration, etc.) and *intuitive convictions* (rooted in a more direct access to moral features, non-inferential evidence concerning ethical value, etc.). Stated in this way, RE is often supposed to establish some sort of compromise between *rationalist* and *sensualist* approaches to ethics: instead of exclusively relying on either reason or perception, both sources are meant to be brought together and to mutually correct each other until some stable trade-off between ethical rationality and moral sense is achieved. Reconstructed in such a manner, RE is thought to be an essential ingredient in the reunification of abstract reflection and concrete insight.[5]

There is a historical dissonance in this description. Intuition, at least in the classical tradition of ethical intuitionism, was mostly thought of as a faculty of practical reason, rather than of moral sense. Consequently, the contraposition of theory and intuition does not unambiguously translate into the opposition of rationalism and sensualism, as is sometimes believed. But I will not delve into this problem. It is far more essential to note that the initial assumption—that RE should hold between theory and intuition—is already flawed: this idea is neither supported by Rawls' original account of the concept in "A Theory of Justice" nor apt to ensure its fruitful application.[6]

It should be recorded, first, that in Rawls' contractualist framework *intuitions* reside on *both* levels of the established deductive/inductive connectivity. They concern the alternative modellings of the initial choice situation as well as the principles of justice that would ensue from them. It is often presumed that only the latter can be subject to intuitive appraisal

---

[5] Instances of this interpretation are to be found in encyclopedia entries [Daniels 2011, § 1, § 4.1; Lowe 1995, 753], meta-ethical work [Brandt 1990, 260–261, 264, 267, 272–273; Brandt 1998, 16–21; Ebertz 1993, 197; Herrmann 1998, 104–105; Holmgren 1989, 46–47; Lyons 1975, 145–147; Singer 1974, 47–48; van Willigenburg 1998, 41, 53–54; Verweij 1998, 34], as well as explorations into the use of RE in applied ethics [Arras 2009, 18–19; Collste 1998, 245; DeGrazia 1992, 520–523, 529–530; Gillon 1996, 260; Rutgers 1998, 231–232; van Thiel / van Delden 2010, 188–189, 198]. Besides, it is often perceptible when RE is thought to reconcile the poles of "theory" and "practice", of "scholarship" and "experience", or similar contrasting pairs.

[6] Referring to the historical tradition Rawls defines "intuitionism" quite narrowly as a group of ethical approaches that contain an irreducible plurality of "competing principles of justice" without any priority setting or hierarchical order [ToJ, § 7, 30]. Consequently, it is exactly the introduction of priority rules and lexical orderings that Rawls supposes to mark a successful "reducing", though not "eliminating entirely the reliance on intuitive judgments" [ToJ, § 8, 39]. The fact that such rules and orderings are integral to Rawls' final set of principles attests that his aim is to *elaborate* existing intuitions, instead of countermanding them, and to *integrate* them in a consistent theory, as shall be explained shortly. However, it is crucial to realise that in Rawls the relevant intuitions are not competing convictions on the level of principles but correlating intuitions on *both* levels, the construction of the initial situation *and* the resulting principles of justice, being balanced in RE to deliver a homogeneous theory.

and, indeed, Rawls affirms that intuitions concerning the principles of justice do exist and are to be appreciated as such: there are "judgments about the basic structure of society", as defined by its principles, which we make "*intuitively*" [ToJ, § 4, 17, my emphasis]. These judgements embody a self-contained standard of adequacy, thus qualifying as "considered convictions of justice" from which reasoning may start [ToJ, § 4, 17]. They are to be respected at least as "provisional fixed points" for further discussion [ToJ, § 4, 18]—exactly because of their *intuitive quality*. But analogous remarks hold with respect to the initial situation: the conception of an "original position", devised as an ethical standpoint for interpreting moral relationships, is also "an *intuitive* notion that suggests its own elaboration" [ToJ, § 4, 19, my emphasis]. It brings in its own standards of adequacy, it can be assessed "from a moral point of view" in its own right [ToJ, § 21, 109], and it is subject to "reasonable philosophical conditions" [ToJ, § 4, 19; cf. § 87, 510, § 87, 514]. More clearly, "the philosophically favored interpretation of the initial situation incorporates conditions which it is thought reasonable to impose on the choice of principles" [ToJ, § 20, 104; cf. § 20, 105], and the "naturalness of these conditions" discloses itself to the attentive thinker [ToJ, § 87, 507]—i.e. to someone who is receptive to their *intuitive soundness*.

By contrast, *theory* in Rawls denotes the resulting overall conception *after* RE between the two levels has been accomplished. Once the basic modelling of the initial situation and the ensuing set of political principles have been brought to a stable correspondence, taking into account their respective intuitive merits, the process of balancing has been successful, delivering the desired theory: it is only *after* the proper setting of the "original position" has been established that a "*theory* of justice" is really achieved [ToJ, § 4, 16, my emphasis]. The "mutual support of many considerations", as provided for by RE, is the essential condition for a "*conception* of justice", in the full sense of an elaborated *theory* [ToJ, § 4, 19, my emphasis].

Thus, rather than weighing theory *against* intuition, RE designates a hermeneutical instrument to flesh out *intuitions on both systematic levels* and finally incorporate them into *one coordinated theory*. It is a heuristic tool that guides intuitions concerning *both* the proper design of the *initial situation* and the moral quality of the *ensuing principles*, letting them correct each other in a constant bidirectional trade-off and having them finally converge in a unique and, at least for the time being, stable theory. At that point, the choice situation has turned into a fully-fledged and well-designed version of the "original position", instructed by inductive feedback from the output side. The principles, in turn, now conform to considered judgements, "duly pruned and adjusted" by deductive reasoning from the input level [ToJ, § 4, 18]. In the end, the mutual information and adaptation of intuitions working on both sides has not only led to their subsequent *adjustment* and eventual *coinciding* ("equilibrium"). Rather, it has brought about their refined *understanding* in the light of each *other* ("reflective") [ToJ, § 4, 18]. It is this hermeneutical clarifying, and not just reciprocal tuning, of relevant intuitions that is the purpose of RE. And it is only this specific procedure of mutual clarification that permits us to call its final result a theory in the first place.

(2) A closer look at Rawls' wording underlines the importance of this reciprocal deepening of available intuitions on both levels: in the central passage introducing RE Rawls describes it succinctly as working "from both ends" [ToJ, § 4, 18]. If we take that phrase literally—as

I suppose we should—it implies more than just an alignment of conceptually correlating but inherently independent points; that would be merely working *on* both ends, and it would, at most, amount to some *equilibrium* between those points. Working *from* both ends, by contrast, implies the creation of something new—a path, a bridge, a tunnel—, some conjunction between the two levels, definitely exceeding their mere correspondence arranged for by their deductive/inductive relationship; only this substantial connection deserves the term *reflective* equilibrium in the first place. Being related in this way, the two levels, initial situation and principles derived, may attribute meaning to each other. The presence of each level, as well as the tentative variation of its parameters, may instruct intuitions concerning the other, clarifying its philosophical background, illuminating its normative import, and thus guiding its proper design.

Rawls provides another telling formulation, now concerning the resulting theory of this process: after mutual support from both levels has been obtained everything is supposed to fit together "into one coherent view" [ToJ, § 4, 19]. Again, this formulation precludes all interpretations of RE as establishing merely an adjustment of separate levels; that would deliver, at most, *two correlating* views. To gain *one coherent* view, by contrast, means to finally obtain a perspective in which all parts are profoundly interrelated; they do not only *fit to* each other, by some remote affiliation of deductive/inductive accordance, but they are *fixed by* each other, each one contributing to the other's proper understanding and inherent calibration. In such a homogeneous theory every component derives its significance from the other. And this significance, gained by the presence of the other level, allows for its enlightened determination, instead of its mere accommodation that could never, as such, be unique.

It is rarely examined how this mutual deepening of intuitions and their final integration into one consistent theory actually proceeds within Rawls' contractualist account. I would like to fill this gap in order to demonstrate more closely the way that RE functions in its original presentation. For reasons of convenience I will deal with the two levels, *initial situation* and *principles derived*, subsequently. But I hope it will become apparent that their establishment and adjustment is really a two-sided enterprise of reciprocal enlightenment on *equal terms* and in *alternate progress*.

(i) Undoubtedly, there are intuitions concerning the adequate modelling of the initial situation. At first sight, these intuitions seem to refer exclusively to some basic *epistemological demands:* Rawls says that the balancing procedure should start from "widely accepted but weak premises", from "generally shared and preferably weak conditions" concerning the initial situation [ToJ, § 4, 16, § 4, 18]. This might be interpreted as evoking certain kinds of quasi-mathematical standards, such as economy or simplicity, symmetry or weakness, that should guide the description of the initial situation and, particularly, the thickness of the veil of ignorance. Intuitions of this kind, however, are yet uneducated by the normative function of the model, i.e. its role in determining political standards for a *just society*. Thus, they have no major business in the process of RE.[7]

[7] Consequently, Rawls explicitly rejects the idea of choosing the "weakest set of conditions to characterize the initial situation", for quasi-mathematical reasons, as "there *is* no weakest set" that might be singled out in this way [ToJ, § 87, 510, my emphasis]. Instead, the essential conditions of the initial situation, and par-

(ii) When taking into account the normative horizon of the initial situation, i.e. its purpose of ultimately delivering *political principles*, the modelling of this situation attains a highly characteristic meaning, open to intuitive appraisal of a much more definite kind: the veil of ignorance now reveals itself as an obvious account of *impartiality*. It is not a neutral mathematical parameter, open to free variation, depending, at most, on some semi-aesthetic preferences of logical elegance. Rather, not knowing one's position in society while selecting society's principles to one's own advantage is a palpable restriction of *procedural fairness* in making political decisions. Its rationale is to hide pieces of information that would otherwise induce straightforward partisanship by self-interested decision-makers. Its objective is to withdraw personal knowledge from the rational participants that they would inevitably use to promote their *selfish interests*.[8]

(iii) At this point it becomes clear how one level of argument—the modelling of the initial situation—obtains its *basic meaning* from the other level of argument—the resulting political principles. And based on that connection, intuitions concerning the initial situation can *be guided* by the other level. They are not isolated convictions of what an adequate initial situation might amount to, eventually to be adjusted to separate intuitions concerning adequate principles. Rather, they gain their very essence from the presence of the other level. They are, unmistakably, intuitions *concerning* the initial situation, but they owe their decisive *content* to the principles that this situation is meant to entail. Being thus equipped, they can further evolve and may legitimately be expected to converge with the principle side to some non-arbitrary, stable point (without, at the same time, being simply adopted from there): when the veil of ignorance is to provide for impartiality within the model, this impartiality is attained to a *higher* degree the *thicker* the veil is. If the purpose of the veil is to preclude information that the participants would use to their own advantage, the *more* information thus precluded the *fairer* the situation is. Concealing the concrete *position* that a participant will be allotted is certainly a minimum requirement of impartiality, when his choice is allowed to be self-interested (apart from being necessary for the model to have a solution, safeguarding unanimity among the participants). But concealing the *probability* of obtaining a certain role is still more impartial, as it gives him even less information to advance his selfish intentions (which are presupposed in order to avoid any additional, uncontrollable normative input into the model, besides the fairness of the choice situation). True, information on probabilities would be *the same* for all participants and, consequently, they would use it *the same way*—e.g. for calculating *identical expected values*. But anyway, they *would use* this information, and they would use it *to optimise their own fate*—exactly what the veil is supposed *to*

---

ticularly the choice of the "veil of ignorance", must be regarded as "more evident *moral* elements of the original position" [ToJ, § 87, 510–511, my emphasis].

[8] It should be noted that Rawls, when stating that natural fortune and social status are "arbitrary from a moral point of view" [ToJ, § 3, 14], or "irrelevant from the standpoint of justice" [ToJ, § 4, 17], is regularly referring to normative standards which are meant to determine his *fictitious model*—not the *real world*. For the most part, his point is that corresponding advantages should have no "use" in making a social "contract" [ToJ, § 3, 14], i.e. in the virtual "choice of principles" in the "initial situation" [ToJ, § 4, 16], for reasons of *procedural impartiality*—rather than that corresponding differences should be eradicated in the actuality of human life, for reasons of *structural equality*. Admittedly, there are passages in which Rawls considers various social arrangements, including his own principles, with regard to their ability to eliminate the influence of the "natural lottery" and of "social contingencies" on distributive shares in reality, too [ToJ, § 12, 64]. But he emphasises that none of these considerations have the status of "an argument" in the strict sense, since in contract theory all arguments "are to be made in terms of what it would be rational to agree to in the original position", properly defined and suitably balanced in RE [ToJ, § 12, 65].

prevent. So impartiality, when established against the background of rational self-interest, not only implies the preclusion of information about one's own future position (even if everybody was given this individual information on his personal fate). It also suggests hiding information on one's own chances of occupying certain roles (even if these chances happen to be identical for all). Consequently, concealing even probabilities seems essential to warrant a maximally impartial choice, as defined within the model. This speaks in favour of the initial situation being designed as a choice under *uncertainty*, rather than *risk*.

(iv) Finally, once it is accepted that decisions under uncertainty should be guided by the *maximin rule*, this calibration on the model side (derived, as stated, from its relation to the principle side) becomes even more specific. Attention is now focused on the worst position in each social reality considered and the system with the best of these worst outcomes is selected. Consequently, the *difference principle* on the principle level is implied, thus completing the deductive line of reasoning. But again, this constellation may also be regarded as confirmed by the principle level, in an inductive sense. For there, an analogous concentration on the worst-off may obtain, now for explicitly *moral* reasons (see below). This outcome supports the suggested *rational* focus on the worst-off, establishing a concrete and robust RE between both layers.

Conversely, the level of principles is informed by the level of model. That is, principles may not only be deduced from different versions of the initial situation. Rather, the *inherent standards* of the principles are affected by the *conceptual background* of the model. In order to highlight the tight correspondence to the reverse influence sketched above I will work through this *analogous process* in largely *parallel steps*.

(i) Obviously, there are intuitions concerning the basic principles of a just society, too. At first glance, these intuitions seem to be restricted to very basic tenets of *political equity:* as examples for suitable starting points Rawls mentions firm and proven convictions that "religious intolerance" and "racial discrimination" are unacceptable [ToJ, § 4, 17]. These convictions may be regarded as elementary expressions of formal justice, based on precepts such as differentiating between persons only for relevant reasons, avoiding arbitrariness in social relations, giving everybody his due, or treating like cases alike. However, intuitions supporting these formal maxims, apart from being consistent with a plethora of concrete arrangements concerning liberties and goods, are not positively referring to the basic contractualist idea, i.e. to the concept of deriving just principles from a *fictitious choice*. Thus, they play no relevant role in the process of RE.[9]

(ii) Once the argumentative groundwork of Rawls' contractualism, i.e. its reliance on a rational *decision procedure*, is taken into account the principle level adopts a much more peculiar flavour and intuitions pertaining to it are directed in a much more illuminating way: there is an unambiguous element of *solidarity* entering the discussion of principles as soon as they are perceived against the counterfactual background of resulting from a self-interested

[9] Of course, the choice model is *perfectly able* to preclude formal injustices of "racial and sexual discrimination", because choosing such principles under a veil of ignorance would be highly "irrational", considering that one could easily turn out to be affected by the disadvantaging practices in question [ToJ, § 25, 129]. Thus, it can be shown, by way of "inference from the theory", that these principles are "no moral conceptions at all, but simply means of suppression" [ToJ, § 25, 130]. My point, however, is that intuitions on the principle level may become much *more concrete*, when considering their counterfactual origin in self-interested choice. This I shall demonstrate shortly.

choice under a veil of ignorance. Just as the fictitious agents in the initial situation are *concerned* with their own fates, a real society built on their principles will be *concerned* with the fates of its citizens, instead of restricting itself to the abstract demarcation of immunities and privileges. Essentially, generating principles for just societies from a rational model of self-interested participants furnishes those principles with the aura of *caring for others* as one would *care for oneself*.[10]

(iii) Now it is the level of principles that receives its *specific significance*—caring about citizens' fates—from the level of the model—caring about one's own fate. And again, this peculiar horizon may help *to instruct* intuitions working on the principle level. These are not left to themselves. They do not have to assess freely varying suggestions of what a just society might amount to, finally having to be adapted to the independent choices of fictitious participants in the initial situation. Rather, they may develop an exceptional standard, *referring*, surely, to *all kinds* of principles that may be proposed for just societies, but *judging* them from the *unique perspective* of making a choice for others as one would choose for oneself. It is not too daring to suppose that this perspective induces certain restrictions on the principle side that may well accord to corresponding qualifications on the model side (without, at the same time, simply importing its results): justice, on account of choosing for others as if one's own fate was at stake, is basically construed by the imagination of *taking on another's part*. Society, corresponding to the idea of caring for citizens the way that rational beings care for themselves, is fundamentally evaluated by contemplating how it would be to *stand in someone else's shoes*. However, in this perspective, *numbers* of positions of a given type will not enter into one's considerations. For those numbers are not essential ingredients of the positions that one steps into, but only external features to them. Taking over someone else's part, if successful, opens the specific viewpoint of that *individual*. And this viewpoint, in its individual nature, is unaffected by the size of the group that he or she may belong to. Thus, exhibiting solidarity with other persons, their natural lot, their social status, by imagining that one might have ended up in their position suggests that it is of no moral concern how many individuals share that kind of role. Solidarity, rooted in the idea of taking on another's part, makes it an irrelevant fact how many people fare similarly. This exclusion of numbers is by far not self-understood in political ethics. *Most* accounts of justice in fact *do* refer to the size of groups (not only utilitarian, but egalitarian ones, for instance, too, depending on their measuring units for social equality). *Solidarity*, however, in the sense established, does *not*. Envisaging someone else's position as one's own remains invariant with respect to numbers (as numbers are not an inherent characteristic of that position). Consequently, political principles, when judged against the notion of making arrangements for others as one would want them to be for oneself, are fully fair only when they disregard the numerical weight of social groups. Solidarity, so it seems, does not care for *frequencies*, but only for *situations*.

---

[10] The *first* element of the choice model, i.e. the "assumption of mutually disinterested motivation", "asks little of the parties", thus lagging behind any *concept of solidarity* and confining society to the mere settlement of "conflicts" [ToJ, § 87, 510]. But once the *second* element, i.e. the "veil of ignorance", requires us "to go beyond a concern for our own interests", it is precisely this element of "concern" that endures and extends to other people's interests, thus giving the impartial choice of principles a definite air of *caring for others* [ToJ, § 87, 511].

(iv) Finally, justice conceptualised as caring about people's fates (induced, as demonstrated, by its reliance on the choice model) may even imply a more definite focus, concentrating on the worst-off, as proposed by the *difference principle*, instead of considering all positions as equivalent. Solidarity, in its own right, displays a noticeable asymmetry in concern, giving priority to the poorest rather than regarding all positions equally. It thus supports the model side in its rational tendency towards the *maximin rule*, along the inductive way of argument. But of course, it is also entailed by the model side, in the deductive direction. The *careful avoidance* of the worst outcome for oneself (see above) translates into *solidary care* for the least advantaged in society. Thus, the former adds content and conclusiveness to the latter, uniting both perspectives, *rational* and *moral*, in a precise and stable RE.

This is but a rough sketch of how Rawls' framework may deepen intuitions on both levels of argument, substantially *directing* them to each other without simply *aligning* them to each other, finally delivering a homogenous theory. Not knowing one's one fate, not even the probabilities of outcomes, may be *most impartial*, taking into account the *political dimension* of the thought experiment. Vice versa, considering every role in society, no matter how many positions it consists of, may be *maximally fair*, accepting the basic idea of *making choices* for others. The conservative maximin rule, securing oneself when probabilities are unknown, may have a *substantial relation*, and not just a *structural parallel*, to the solidary difference principle, optimising the fate of the least advantaged irrespective of how many they may be.

I do not want to claim that this concrete result of Rawls' considerations is incontestable. On the model level, there may be reasons to regard the risk version as altogether sufficient for impartiality and to deny the logic of "the thicker the veil—the more impartial the choice". On the principle level, there may be reasons to hold that numbers of positions in different classes are inherently relevant to the assessment of a society's justice, even on a "solidary standpoint". Additionally, there are decision-theoretic reservations against the maximin rule and ethical caveats against the difference principle, questioning their exclusive focus on the least advantaged instead of considering worst and best outcomes on a par. But whatever the results of these discussions may be, the basic process of Rawls' thinking, its conceptual embedding and reciprocal enrichment, is well-defined and non-arbitrary. Particularly, any alternative suggestion for RE would have to distinguish itself by the same merits of working "from both ends" and integrating its components into "one coherent view". Whatever may be thought about the adequate modelling of the initial situation and proper principles of justice ensuing from it, it must not amount to a mere adjustment of isolated components but establish a co-ordinated deepening and mutual influencing of relevant convictions concerning both the proper setup of the initial situation and the requirements of justice in society. To do so it must argue for a substantial relationship between the two levels, with intuitions on both sides being directed by the presence of the other and finally converging into one counterbalanced theory.

RE in this sense may be a helpful tool for intensified discussions of normative positions and progressive adjustments of their pivotal parameters. But its successive enriching and informed balancing depends crucially on the presence of an overarching argumentative structure that, first, conjoins its various components in a clear-cut *deductive/inductive relationship* and, second, allows them to attribute *specific sense* to each other. It is only in this way

that pre-existent intuitions may be mutually adapted and finally transformed into a homogeneous theory of optimised elements, as outlined above. By contrast, the idea of balancing *theories* against *intuitions* lacks exactly this common framework which might make any adjustment procedure between its components a credible enterprise. Theories and intuitions, taken as independent approaches to moral phenomena, are not embedded in any higher structure *embracing* both of them and *relating* them in an informative way. Without such a uniting structure, however, they do not stand in a deductive/inductive relationship and cannot contribute substantially to each other's interpretation.

(3) This being said, Rawls' approach can be defended against common accusations that are based on a deficient understanding of RE and, particularly, the interplay of theory and intuition that it establishes. An exemplary reference is an early critique of Rawls' "A Theory of Justice" by Richard Hare who reproaches him emphatically for presenting a philosophical conception that eventually breaks down to some shallow and badly concealed intuitionism, attempting to tailor a suitable theory to predetermined intuitions.

Hare stresses that there are alternative modellings of the initial situation, primarily Harsanyi's risk version and Rawls' uncertainty version, leading to different results on the principle side, i.e. utilitarianism and the difference principle, respectively. For the most part, however, Hare considers intuitions to reside only on the outcome level (i.e. the alternative principles of justice), whereas he believes the model level to be intuitively neutral (constituting optional alternatives of, so to say, "pure theory"). Consequently, the selection of a specific model and its corresponding outcome appears to be a one-sided matching of a free-to-choose theory to pre-existent intuitions: in Rawls, Hare assumes, "the theoretical structure is tailored at every point to fit Rawls' intuitions" [Hare 1973, 84]. Preferring the difference principle for intuitive reasons Rawls selects the compatible theory without any deeper justification than his own bias towards the result. Someone else, advocating utilitarianism for contrary intuitions, might just as well pick the risk version of the initial situation, thus deriving his preferred result of average utility maximisation. Rawls, according to Hare, is making "constant appeal to intuition instead of argument", drafting "his theory to suit his anti-utilitarian preconceptions" [Hare 1973, 94–95].

Interpreted in this unidirectional way, RE could hardly be defended against collapsing into pure intuitionism of a very primitive kind. It would not contribute to any deepening of relevant intuitions but be restricted to the purposeful selection of a fitting theory. Instead of paving the way to an objectively stable state of balance, it would split up into a plurality of highly instable equilibria without clarifying in the least which of them should be chosen. However, this disappointing result depends crucially on an essential primary misconception: it is based on the idea that RE should be an adjustment between theory (alternative designs with no inherent moral valence, thus open to any external parameter setting) and intuition (moral default convictions insusceptible to further instruction, thus remaining fixed in their original course). In this case, there is indeed no way to avoid the conclusion that one is free to pick any theory suited to deliver the result one's intuition prefers. However, once it is conceded that there are moral intuitions operating on the model level, too, one may aspire to balance basically akin parts (moral intuitions on the model level against moral intuitions on the principle level) and, by doing so, to eventually obtain something

new (a deepened understanding and substantial directing of these intuitions, finally converging into a coherent theory).

At some points, in fact, Hare seems to allow for intuitions working on the model side as well: in advocating Harsanyi's conception he states that the risk version, being sufficient to secure impartiality, distinguishes itself by applying a "very economical veil" [Hare 1973, 90; cf. ibid., 93, 101]. Banning only a minimum of information, it allegedly delivers "the simplest form" that the "rational contractor theory" can possibly take [Hare 1973, 95; cf. ibid., 91, 101]. I leave aside whether the risk version is indeed more "economical" (the veil may be thinner, but, consequently, the number of variables to be dealt with is larger) or comprehensibly "simpler" than the uncertainty version (it may be easier to formulate, but, presumably, its solutions are more contentious). At any rate, these qualities of economy and simplicity, even if they were uniquely attributable, are morally irrelevant. They may be conceptually attractive, representing some of the above-mentioned quasi-mathematical virtues (i) that choice models can display. But they remain unaffected by any deepened understanding of what the whole construction is devised for.

Thus, economy and simplicity, as exposed by Hare, may seem preferable to exactly that type of intuition which is confined to the model level and unaware of the principle level. They appeal to appraisals of the choice situation prior to any kind of mutual information and balancing within the argument—i.e., prior to any relevant exchange, let alone substantial RE between the two levels. But then, it is certainly not these intuitions that should tip the scales between the risk and the uncertainty version. They are nothing but some preliminary, and eventually dispensable, attitudes concerning mathematical convenience—way before the hermeneutical project of clarifying the meaning and guiding the direction of intuitions on both levels by relating them to each other has finished, or even begun. Once the moral dimension of impartiality is acknowledged on the model side by considering its purpose to deliver principles of justice, once the tangible idea of solidarity enters the principle side by realising its origin in a choice model, intuitions concerning economy or simplicity present themselves as idle side aspects in determining the proper parameters of the thought experiment. It is only after intuitions on both sides have been deepened (working "from both ends") and theory has been established by integrating them (into "one coherent view") that both model and principles gain their relevance and justification. Economy and simplicity, as merely mathematical merits, have no part in that.

At times it appears that Rawls himself tries to resort to purely mathematical standards of type (i) in order to argue for his uncertainty model: I already quoted a passage in which he suggests starting from "preferably weak conditions" for the initial situation [ToJ, § 4, 18]. Later on he states that the reasons for the "veil of ignorance" go beyond, but hence include, "simplicity" [ToJ, § 24, 122]. Again, I skip the question of which version of the initial situation, risk or uncertainty, is "weaker" or "simpler". One might suppose that Rawls' idea of excluding all particular knowledge of the future society (such as knowledge of the numbers of individual positions in the different roles) while admitting all general knowledge (especially knowledge of psychological, sociological and economic laws which enable to anticipate the societal effects of various principles considered) is meant exactly to embody such conceptual weakness or simplicity [ToJ, § 24, 118–120]. But the opposite stance, re-

garding the admittance of all knowledge concerning the future society as conceptually weaker and simpler (though it leaves more data to deal with) than drawing a line between those two types of knowledge (which might turn out to be harder than expected), is certainly just as comprehensible. Anyway, it seems that Rawls, just like Hare, tries to resort to pseudo-mathematical standards of conceptual ease in order to justify his model, in spite of their demonstrated infertility in RE.

However, a closer look at the passages in question reveals that, actually, they are not at all meant to determine the precise thickness of the "veil of ignorance". As opposed to Hare who exploits the standards of economy and simplicity in order to argue for the risk version, Rawls regards weakness just as a very preliminary starting point for further elaboration of what "reasonable conditions" on the model side might eventually look like [ToJ, § 4, 18], while simplicity is introduced as only one of the "reasons for the veil of ignorance" [ToJ, § 24, 122]—i.e. as an argument for using just *any* veil, not the *specific* veil that Rawls eventually endorses. To be more precise, simplicity in the given context refers to the solvability of the model: "The veil of ignorance makes possible a unanimous choice of a particular conception of justice. Without these limitations on knowledge the bargaining problem of the original position would be hopelessly complicated" [ToJ, § 24, 121]. That is, if there were no restrictions of knowledge in the initial situation, the participants would certainly not agree, as everyone would be opting for those principles that serve his individual purposes best. It is only by concealing information on their individual positions in society that they find themselves in the same situation of choice and thus, being equally rational, will select the same principles of justice. This is what Rawls calls "simplicity": the "veil of ignorance" nullifies the effects of "specific contingencies which put men at odds" and thus preclude the participants from finding a common solution [ToJ, § 24, 118]. Consequently, it avoids the conceptually intricate and ethically dubious task of reconstructing some kind of net preference or majority vote from their divergent decisions. Instead, it allows for "unanimity" in their selection of principles, which, taking into account the moral significance of the model, may be interpreted as representing "a genuine reconciliation of interests" [ToJ, § 24, 122]. But obviously, this kind of "simplicity", in the sense of enabling agreement, is warranted by any veil of ignorance refusing personal information—Rawls' as well as Harsanyi's.

A parallel remark holds for another statement concerning the veil of ignorance that Rawls makes in the same paragraph. The whole passage referred to above reads: "Now the reasons for the veil of ignorance go beyond mere simplicity. We want to define the original position so that we get the desired solution" [ToJ, § 24, 122]. This (admittedly unfortunate) phrasing (designing the "original position" in order to obtain the "desired solution") has been readily picked up by Rawls' critics, including Hare, as a clear confession of one-sided intuitionism. Allegedly, it demonstrates that Rawls unabashedly forges his "theory" in order to suit the "conclusions he wants to reach" for merely intuitive reasons [Hare 1973, 91]. It frankly declares that the denial of probability information, in itself "quite arbitrary" as a constraint on the model, is only introduced in order to obtain "conclusions which he finds acceptable", on the level of principles [Hare 1973, 104].

However, this supposedly treacherous phrase about the "desired solution" is located, as the quote reveals, immediately after Rawls' considerations of "simplicity". And so, it must be read in the same argumentative context: Rawls is still talking about the reasons for introducing *any* veil of ignorance—he is not yet talking about the *thickness* of the veil, i.e. about the problem of concealing probabilities or not, as Hare suggests. Of course, when appealing to the "desired solution" Rawls advances *further reasons* for the veil of ignorance (reasons that "go beyond mere simplicity")—but these new reasons concern the *same question*, i.e. the question of why applying any veil at all (the whole passage is about "the reasons for the veil of ignorance" as such). And in specifying these new reasons Rawls is quite explicit that they cannot be cited to argue for the uncertainty version as opposed to the risk version. The purpose is now to prevent the outcome of the model from being "biased by arbitrary contingencies", i.e. to exclude the "arbitrariness of the world" from the initial situation [ToJ, § 24, 122]. But, obviously, this is nothing but a restatement that the systematic function of the veil of ignorance is to introduce impartiality into the model. It is meant to prevent the participants from exploiting "social and natural circumstances to their own advantage" [ToJ, § 24, 118]. It is supposed to keep them from knowing "how the various alternatives will affect their own particular case" [ToJ, § 24, 118]. This purpose, however, is fulfilled by any veil, as long as it precludes information on the participants' individual positions in society. So, in inviting us "to define the original position" in a way to "get the desired solution", Rawls is not aiming at *a specific solution*, such as the difference principle, that we may "desire" for some precise intuitions on the *principle level* and for which we should select the appropriate theory (by choosing a *thick* veil). Rather, he is searching for essential ingredients apt to deliver *some impartial solution* which we "desire" for very basic intuitions on the *model level* and which we may only later, relying on RE, elaborate into a theory (but provisionally requiring no more than just *any* veil).

IV. Balancing general principles against particular judgements?

(1) Another widespread conception of RE, often closely associated with the previous one, assumes that it should hold between *general principles* (ethical rules, moral laws, possibly provided for by theory) and *particular judgements* (referring to singular cases, individual situations, potentially relying on intuition). Understood in this manner, RE allegedly offers some kind of reconciliation between *generalist* and *particularist* accounts of ethics: instead of relying exclusively on either universal norms or singular prescriptions, RE unites both aspects and integrates their respective qualities in order to achieve the most adequate balance between them. Especially, it precludes a cheap understanding of applied ethics as simply "applying" given norms to concrete cases by "subsuming" individual instances under covering laws.[11]

[11] This second perception of RE prevails in a considerable amount of literature, including handbook articles [Blackburn 2005, 312; Buchanan 1992, 657; Daniels 2011, § 1, § 3.1, § 3.2.1, § 4.1, § 5; DePaul 2006, 599–600, 602–604; Dworkin 2006, 639; Lowe 1995, 753; Scanlon 2003, 140–141, 150–151; Simon 1992, 1163; Solomon 1995, 821], meta-ethical considerations [Brandt 1990, 260–261, 266–267; Brandt 1998, 16–21; Daniels 1996, 1–3, 6–8, 22, 27–28, 30–31, 334–340; DePaul 1993, 13, 16, 19; Ebertz 1993, 194, 197, 202–203; Herrmann 1998, 104–105; Holmgren 1989, 43; Lyons 1975, 145–147; Nagel 1973, 221; Raz 1982, 307–308, 317–318; Singer 1974, 47–48; Tersman 1993, 19, 122–123; van der Burg / van Willigenburg

There is a systematic problem with the idea of settling the classical opposition between generalism and particularism by squaring the deductive line ("from principle to judgement") with the inductive line ("from judgement to principle"). The reason is that, at least in one essential meaning of the word, a particularist denies the very existence of general principles and regards morality as exclusively constituted in particular judgements. Correspondingly, she does not want to work inductively, from particular cases to general principles ("bottom-up"), as clearly as she would not allow for the reverse line, from general rules to particular judgements ("top-down"). Both pathways, the deductive as well as the inductive, presuppose the generalist commitment that moral principles do exist and logically govern the validity of moral judgements. The adherents to those pathways, deductivists and inductivists, merely disagree on the question of how to best obtain those principles. Some want to establish them directly, in the form of "axioms" (like in mathematics), others try to derive them indirectly, by conclusion from "observations" (like in the natural sciences). A particularist, in the strict sense of the word, moves in neither direction, holding that there are no principles at all in ethics, but only collections of cases. This attitude precludes both ways, no matter whether they should start from principles (deductively) or whether they should lead to principles (inductively). Correspondingly, RE could certainly not reconcile generalism and particularism in the way envisaged, but only deductivism and inductivism. However, I shall not dig any deeper here, as the matter is, for the most part, terminological in nature. The essential point is that the underlying conception of RE—balancing general principles vs. particular judgements—is already incorrect: it has no basis in Rawls' account of RE in "A Theory of Justice" and it does not, in any case, represent a promising line of inquiry.[12]

(2) It is easy to see that in Rawls' central argument particular judgements do not enter at all. The whole balancing procedure takes place between the *fundamental modelling* of the initial situation and the *general principles* of justice that derive from it. Neither level is concerned with the evaluation of singular cases. The former envisages a fictitious thought experiment that has no status as a real event, the latter represents the most general rules that a society can possibly adopt.

Strictly speaking, there is not even a comprehensible difference in generality whatsoever between these two layers of argument. Of course, they do stand in a *deductive/inductive rela-*

1998b, 1–2, 14; van Willigenburg 1998, 41, 47–54; Verweij 1998, 31–34], as well as investigations into the deployment of RE in applied ethics [Arras 2009, 18–19; Beauchamp / Childress 2009, 381–382, 384; Collste 1998, 245; DeGrazia 1992, 520–523, 529–530; Ebbesen / Pedersen 2007, 38, 42; Rutgers 1998, 231–232; Strong 2010, 126–127, 135–136; van den Beld 1998, 73–74; van Thiel / van Delden 2010, 188–189, 195; Widdershoven 2007, 50–52]. Basically, wordings that RE should dissolve confrontations between "principles" and "judgements", between "norms" and "convictions", etc., are often in line with this interpretation.

12 In fact, there seems to be only one instance in "A Theory of Justice" where Rawls notes that "particular cases" might lead to a revision of "our judgments" [ToJ, § 4, 18]. But this isolated remark is at variance with Rawls' basic conception and actual use of RE throughout the book which is exclusively concerned with the matching of conditions for the (hypothetical) initial situation and judgements on the (universal) principles derived from it. Consequently, incidental passages in which Rawls contrasts "general conceptions or particular convictions", or "[f]irst principles and particular judgments", should be received with care [ToJ, § 4, 19, § 87, 507]: sometimes Rawls is not explaining *his own* understanding of RE, but rather criticising *other traditions* of balancing and judging, sometimes the two components do not refer to convictions concerning universal rules and convictions concerning *singular cases*, respectively, but rather to convictions concerning the initial situation and convictions concerning *specific principles*, devoted to certain realms of society, special fields of justice, etc.

tionship. But this relationship is not based on the concept of *application and subsumption*. Model parameters and principles deduced are not connected in the way of leading from the *general* to the *particular*, like rules and cases in a simple syllogism. Rather, they are related by a decision-theoretic groundwork that suggests certain *choice rules* for specific *situation types*, and by a contractualist thought experiment that translates these *decision rules* into *distributive principles*. It is hard to see why the model level should count as "more general", the principle level as "more particular", or possibly the other way round, although parameters of the former imply results in the latter, and outcomes in the latter suggest the setup of the former.

At the same time, it seems to be precisely this connection of components, affiliated not by syllogistic relations of application and subsumption, but by conceptual ties of decision theory and contractualist reasoning, that allows for the mutual, coordinated deepening of intuitions on both levels and their integration into one coherent theory, as demonstrated above. By contrast, *general principles* and *particular judgements*, though straightforwardly related by way of logical deduction/induction, do not share this kind of essential cross-referring that could make their balancing more than an adjustment of isolated parts. There is no *substantial connection* between rules and cases, as there is between the modelling of a choice situation and the principles of justice ensuing from it (via the decision standards that are implied by the modelling and the contract view that translates them into principles); there is just a *simple assigning* of cases to rules, as the latter trivially embrace the former. There is no *common balancing* of rules and cases, as there is for the impartial formulation of the initial situation and the solidary account of political principles (because each of them is informed by the background of the other); there is just a *separate alignment* of rules and cases, as the latter evidently falsify the former.

(3) I may have an intuition that *lying in general* is wrong, and another intuition that *a particular lie* is wrong. They may seem "close" as they are somewhat directly related, by way of logical or, if you prefer, deontological implication. But there is no *substantial conjunction* between them that might allow for any deepening elaboration, mutual information, etc. There is only their *logical relation*, which can hardly contribute to their further concretion into a stable theory. The fact that the rule never to lie trivially demands the condemnation of an instance of lying, or that the rejection of a concrete lie is evidently covered by the abstract prohibition to lie, does not contribute to any better comprehension on either side. Logical subordination does not advance hermeneutical understanding.

By contrast, I may have an intuition concerning the *proper modelling of the initial situation*, and another intuition concerning *decent principles of justice in society*. They are certainly much further "apart" from each other, as they share no immediate relationship of logical entailment. But exactly because of this they can be *conceptually connected* in a construction that establishes a decision-theoretic link and a contractualist association between them. And here, the meaning of the model and the significance of the principles can be *mutually unfolded*. To understand that the veil must ensure impartiality in political decisions, to accept that society should care for its members just like rational beings care for themselves, is a substantial enrichment of normative insights into both levels. Following these lines of mutual information there is hope to attain RE between both layers.

Confronting the proposed rule to never lie with a possibly justified single lie may lead to a weakening of the general principle or to a correction of the particular judgement. It may encourage parties to admit exceptions to their universal norms or to qualify their statements on specific cases, in some sort of spontaneous concession to the other. It may invite them to go back and forth, changing from the general to the particular or vice versa, trying to figure out which side appears more forceful or reliable to them. But it will never trigger that simultaneous, illuminating process of *seeing* one level *in view* of the other. It will never help to *better understand* one side in light of its output, or to find its *justification* with regard to its roots. This is because there is no conceptual framework that might provide for this attribution of significance, or for this consciousness of origin. They *negate or falsify* each other, squarely. But they do not *clarify or direct* each other, hermeneutically.

Thus, trying to establish RE between rules and cases may articulate their plain logical relationship. It may call to the fore competing intuitions on both sides. But it will never create *something new* between them. Their connection is completely exhausted by the formal interplay of logical entailment, leaving no room for any material links that might establish some deeper conjunction between them. You do not better understand a principle in light of a case, you do not better comprehend a judgement in light of a rule. You may specify the rule by learning about its particular consequences, you may generalise the case by detecting its conceptual core, you may change your mind when you let the other level influence your previous opinion. But you do not gain *substantial insight* into what your present opinion is all about. You will not have it pursue its own destination in a way better informed by the other side.

V. Balancing divergent moral positions on the same conceptual level?

(1) A final popular viewpoint assumes that RE should be sought between *competing ethical conceptions* (theories, principles, etc.) or between *rival moral statements* (intuitions, judgements, etc.) on the same level of abstraction or concretion, respectively. Thus conceived, RE is meant to cope with *ethical pluralism* on various levels of moral thinking when a consensual higher-ranking *master position* is not available: it is supposed, e.g., to balance Kantian ethics against utilitarianism, or divergent positions concerning the moral legitimacy of concrete actions. In applied ethics, for instance, it is expected to deliver a justified compromise between the principle of autonomy and the principle of beneficence, or between the moral stances of embryo protection and freedom of research.[13]

This third conception of RE is probably the most ambitious one. RE is now thought to combine ethical positions that are most fundamentally divergent, being based on different anthropological assumptions, laying emphasis on dissimilar normative accounts and making use of unlike philosophical categories. It is possibly a matter of dispute whether a success-

[13] This interpretation of RE, probably constituting the most common strand in its current reception, is less prevalent in scholarly literature, but does show up in basic accounts of the concept [van der Burg / van Willigenburg 1998b, 3–4], as well as in standard portrayals of applied ethics [Gillon 1996, 260]. Anyway, it is more or less straightforwardly included whenever RE is supposed to unite and trade-off an ultimately unrestricted multitude of moral and non-moral positions, like in "wide" RE, so that it is at least implicit in a plurality of the works cited in footnotes 13 and 25 that refer to this idea of "wide" RE.

ful alignment of this sort would still deserve the title "pluralist" in the first place, or whether its arrangement of components would eventually be "monist" in nature. If the latter was the case, there might be an element of contradiction in the attempt to unite divergent ethical positions without wanting to disturb their asserted independent standing. But no matter whether the idea—balancing divergent moral positions on the same conceptual level—is compatible with the precepts of pluralism or not: it is certainly not in accord with Rawls' notion of RE in "A Theory of Justice" or with any reliable mechanism that might be brought under that title.[14]

(2) In Rawls, RE is applied within a unique, overarching conceptual framework that establishes a clear-cut, well-defined deductive/inductive relationship between the components to be balanced: it weighs certain variants of the initial situation against the principles of justice that would follow from them. It is within this framework that the individual character of the alternatives is defined (risk vs. uncertainty, utilitarianism vs. difference principle) and can be described as variations of certain parameters (the thickness of the veil, with its ensuing principles). It is also within this framework that the decisive questions can be asked (what is the proper thickness of the veil? which are the adequate principles of justice?) and finally be answered in a mutual deepening and directing of relevant intuitions (eventually converging into one coherent theory).

This overarching array, this stable setting of alternatives and common basis of reasoning, is missing where fully independent approaches (deriving from different ethical traditions) or largely unrelated statements (of a simple pro/con type) are brought together and supposed to be balanced against each other. They constitute fundamentally incommensurate perspectives (Kantianism vs. utilitarianism), they amount to basically disparate positions (embryo protection vs. freedom of research) the balancing of which would amount to little more than some ad hoc procedure of arbitrary compromise.

Particularly, these conflicting moral theories and convictions do not stand in any kind of deductive/inductive relationship to each other: Kant does not entail or preclude utilitarianism the way the uncertainty version of the initial situation entails the difference principle and precludes utilitarianism. Embryo protection and freedom of research are not connected by straightforward implication or contradiction the way the risk version of the initial situation implies maximisation of average well-being and contradicts focussing on the worst-off. All these different stances are *at variance with* each other, and sometimes their implications clash. But they are *not associated with* each other, like the model level and the principle level in a contractualist account. Consequently, they do not allow for establishing RE in the full sense of a reciprocal balance—occurring along a line of deduction/induction

[14] Indeed, Rawls is willing to complement his own "Kantian" approach [ToJ, § 40, 221–227, § 87, 511] with certain additional tenets, particularly with "efficiency" in the sense of Pareto optimality and with the utilitarian standard of "maximizing the sum of advantages" [ToJ, § 46, 266]. In his second priority rule these are mentioned as subordinate to his second principle of justice which, in turn, is subordinate to the first. However, Rawls regards "efficiency" explicitly not as a requirement of "justice", but as an *additional concern* of social systems [ToJ, § 1, 5, § 12, 60, § 18, 94, § 41, 230], while utilitarianism enters only because Rawls, for reasons of conceptual modesty, assumes that the decision-makers in the "original position" merely *relatively rank*, rather than absolutely distinguish, alternative principles, eventually favouring his solution over the "principle of average utility" and this one, in turn, over the "classical principle" of utility [ToJ, § 26, 130, § 29, 159, § 30, 160]. Consequently, both tenets are not part of his *affirmative* reasoning for principles of *justice*, let alone equitable associates in any kind of RE, but only marginal attachments and minor concessions, transgressing the realm of justice and proving second-best in their argumentative solidity.

that connects the levels in question unambiguously, and emerging from the systematic interdependence and mutual correction of its own constituents. There can never be a substantial conjunction between them, attained by working out intuitions "from both ends", for they *are not* conjoined in any systematic way (like an initial situation and the principles adopted). There can never be a common theory, embedding them in "one coherent view", as they articulate *systematically separated* views on the problem exposed (as opposed to the original position and the difference principle).

The situation is definitely worse than in the antecedent section: *general principles and particular judgements*, at least, share a *logical* relationship of deduction/induction (though this relationship does not allow for any substantial deepening and bidirectional adjusting of the two levels). *Divergent moral positions on the same conceptual level*, however, do not stand in *any* relationship of deduction/induction (let alone in a relationship that might be substantiated by mutual enrichment). Thus, they certainly do not qualify for RE in any relevant sense. There can never be an "equilibrium" between parts that are not connected in some way—physically, mechanically, logically, etc.; for *equilibrium* is a state of balance between bodies, particles, propositions, or any entities that are somehow *acting* on each other. All the more, there cannot be a "reflective" equilibrium between philosophical positions that do not even share a common framework, but simply designate opposing traditions and convictions; for *reflective* equilibrium requires a stable *connection* between the levels to be balanced so that their eventual elaboration is reproducibly guided by the conceptual interplay and mutual adaptation of these levels.

(3) The second notion of RE deprived the idea of its conceptual stringency, failing to arrange for a substantial conjunction between its components (general principles and particular judgements). But at least, it respected the basic requirement of some connection between the levels in question (the logical link of application and subsumption). This third understanding of RE lacks even this kind of elementary relationship. Consequently, it is doomed to result in the most arbitrary compilations of highly heterogeneous components, following maxims like: "Let's mix some categorical imperative with some utility maximisation, because both have a say in ethics, in some way, to some degree, and we do not know, or we cannot agree, which one to prefer." Ethical chimeras of this sort represent the exact opposite of what RE is meant to attain. They are by no means "reflected", in the sense of a deepened understanding by mutual instruction, and they are not even "equilibria", in the sense of some counterbalancing of related components. All they constitute are superficial concessions to factual disagreement.

So mixing "some Kant" with "some Mill", "some embryo protection" with "some freedom of research", without an embracing framework that assigns them their systematic place and relates them in a mutually instructive manner, is not a promising enterprise. Particularly, it is deplorably remote from what Rawls calls RE, and from what RE can actually mean. Sure enough, *Rawls uses* RE to *decide between* competing convictions of justice that range on the same level of abstraction, such as utilitarianism, liberalism, his own set of principles, and others. But *RE itself* does not *hold between* those rivalling conceptions. Rather, it holds between *just one* of these conceptions and *its corresponding* modelling of the initial situation. And

the fact that exactly *these two* stand in that equilibrium, and not one of the *other pairs* of principles and modelling, is what distinguishes them and gives them their ethical justification.
Thus, RE does not stand "vertically", between *competing* normative stances (e.g. rival principles for political societies), trading off their opposing attitudes and assembling them into some sort of compromise. Rather, it stands "horizontally", between *corresponding* normative stances (i.e. various models and their respective principles), helping to adjust their correlating parameters and finally selecting their best choice. To search for RE between Kant and Mill, or between embryo protection and freedom of research, is, as it stands, an utterly absurd enterprise. It will end up in some shallow, merely intuitive blend of principles and convictions without any theoretical background that might make this procedure more than an ad hoc affair of philosophical bargaining. Rawls does not balance doctrines or convictions against each other, but possible modellings and ensuing principles. For these are the two sides of a conceptual unity, and, as such, may exert a concerted, well-directed normative impact on each other.

VI. Conclusion

One may question the basic ethical cogency of Rawls' contractualist approach. One may have doubts concerning the rational uniqueness of his decision-theoretic assumptions. One may argue that RE should mark out a different situation of choice as the "original position" and, correspondingly, present a different set of principles as proper realisations of "justice as fairness". Be that as it may, one has to admit that Rawls' conceptual framework brings together different layers of normative reasoning in a way that allows for their mutual adjustment in a well-defined, fruitful and conclusive procedure. Occasional amendments to his argument may be in place. But the basic setup of this argument is non-arbitrary, fertile, and convincing.
It was demonstrated that each of the above-mentioned interpretations considerably departs from Rawls' original conception of RE in "A Theory of Justice": it is not *theory* and *intuition* that are weighed against each other, for intuition is the primary material on both levels and theory is the final result of the completed process. *Particular judgements* never enter the thought experiment, so they cannot be balanced against *general principles*. *Divergent moral positions* on the *same conceptual level* are not compared either, as they do not stand in the decision-theoretic, contractualist connectivity that Rawls exploits. I also argued that these alternative understandings of RE deprive the concept of its soundness: the components they try to align are incommensurate in their epistemological standing (theories and intuitions), trivially related by logical subordination (general principles and particular judgements), or devoid of any conceptual relationship (rival ethical approaches or moral statements). Consequently, in each case they fail to unite the respective components in an argumentative structure that is sufficiently rich to provide for their mutual enlightenment and correlated adjustment. But this is what RE must do, if it is to be an expedient instrument in ethical reasoning.

In summary, it appears that RE is a concept owing its reliability to a well-defined systematic framework in which it is applied (such as Rawls' contractualist approach). However, this framework presupposes certain normative commitments (e.g. acknowledging the moral significance of a rational decision process in a fair choice situation for deriving political principles). Consequently, RE is not a neutral tool for deciding between just any rival moral conceptions, however different and unrelated they may be. Rather, it presumes a definite context of normative premises in order to do its work on the corresponding subordinate components. This restriction may deprive it of the wide range of applications that contemporary ethics tends to ascribe to it, e.g. in applied ethics, and also of the bold promise to provide justification independent of any foregone foundation, as supposed by ethical coherentism. But it seems that this is the price to pay for its conceptual validity.

References

Allais, M. (1953): "Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école americaine". Econometrica 21 (4), 503–546.

Arras, J. (2009): "The Hedgehog and the Borg: Common Morality in Bioethics". Theoretical Medicine and Bioethics 30 (1), 11–30.

Barber, B.R. (1975): "Justifying Justice: Problems of Psychology, Politics and Measurement in Rawls", in: Daniels 1975, 292–318.

Beauchamp, T.L. / Childress, J.F. (2009): Principles of Biomedical Ethics, 6th ed., New York / Oxford: Oxford University Press.

Blackburn, S. (2005): "Reflective Equilibrium", in: Blackburn, S.: The Oxford Dictionary of Philosophy, 2nd ed., Oxford: Oxford University Press, 312.

Brandt, R.B. (1990): "The Science of Man and Wide Reflective Equilibrium". Ethics 100 (2), 259–278.

Brandt, R.B. (1998): A Theory of the Good and the Right, 2nd ed., Amherst: Prometheus Books.

Buchanan, A. (1992): "Justice, Distributive", in: Becker, L.C. / Becker, C.B. (eds.): Encyclopedia of Ethics, Vol. 1, Chicago / London: St. James Press, 655–661.

Collste, G. (1998): "Infanticide in Reflective Equilibrium?", in: van der Burg / van Willigenburg 1998a, 239–250.

Daniels, N. (1975) (ed.): Reading Rawls. Critical Studies on Rawls' *A Theory of Justice*, New York: Basic Books.

Daniels, N. (1996): Justice and Justification. Reflective Equilibrium in Theory and Practice, Cambridge: Cambridge University Press.

Daniels, N. (2011): "Reflective Equilibrium", in: Zalta, E.N. (ed.): The Stanford Encyclopedia of Philosophy (Spring 2011 Edition), http://plato.stanford.edu/archives/spr2011/entries/reflective-equilibrium.

DeGrazia, D. (1992): "Moving Forward in Bioethical Theory: Theories, Cases, and Specified Principlism". The Journal of Medicine and Philosophy 17 (5), 511–539.

DePaul, M.R. (1993): Balance and Refinement. Beyond Coherence Methods of Moral Inquiry, London / New York: Routledge.

DePaul, M.R. (2006): "Intuitions in Moral Inquiry", in: Copp, D. (ed.): The Oxford Handbook of Ethical Theory, Oxford: Oxford University Press, 595–623.

Dworkin, G. (2006): "Theory, Practice, and Moral Reasoning", in: Copp, D. (ed.): The Oxford Handbook of Ethical Theory, Oxford: Oxford University Press, 624–644.

Dworkin, R. (2002): Sovereign Virtue. The Theory and Practice of Equality, Cambridge (Massachusetts) / London: Harvard University Press.

Ebbesen, M. / Pedersen, B.D. (2007): "Using Empirical Research to Formulate Normative Ethical Principles in Biomedicine". Medicine, Health Care and Philosophy 10 (1), 33–48.

Ebertz, R.B. (1993): "Is Reflective Equilibrium a Coherentist Model?" Canadian Journal of Philosophy 23 (2), 193–214.

Gillon, R. (1996): "Ethnography, Medical Practice and Moral Reflective Equilibrium". Journal of Medical Ethics 22 (5), 259–260.

Goodman, N. (1954): Fact, Fiction, and Forecast, London: The Athlone Press.

Hare, R.M. (1973): "Rawls' Theory of Justice", in: Daniels 1975, 81–107.

Harsanyi, J.C. (1953): "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking". The Journal of Political Economy 61 (5), 434–435.

Harsanyi, J.C. (1975): "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory", in: Harsanyi, J.C.: Essays on Ethics, Social Behavior, and Scientific Explanation, Dordrecht / Boston: D. Reidel Publishing Company 1976, 37–63.

Harsanyi, J.C. (1982): "Morality and the Theory of Rational Behaviour", in: Sen, A.K., Williams, B.A.O. (eds.): Utilitarianism and Beyond, Cambridge: Cambridge University Press, 39–62.

Herrmann, E. (1998): "Rationality, Warrant and Reflective Equilibrium", in: van der Burg / van Willigenburg 1998a, 103–114.

Holmgren, M. (1989): "The Wide and Narrow of Reflective Equilibrium". Canadian Journal of Philosophy 19 (1), 43–60.

Hübner, D. (2001): Entscheidung und Geschichte. Rationale Prinzipien, narrative Strukturen und ein Streit in der Ökologischen Ethik, Freiburg i.Br. / München: Karl Alber.

Hübner, D. (2009): Die Bilder der Gerechtigkeit. Zur Metaphorik des Verteilens, Paderborn: mentis.

Lowe, E.J. (1995): "Reflective Equilibrium", in: Honderich, T. (ed.): The Oxford Companion to Philosophy, Oxford / New York: Oxford University Press, 753.

Lyons, D. (1975): "Nature and Soundness of the Contract and Coherence Arguments", in: Daniels 1975, 141–167.

Nagel, T. (1973): "Rawls on Justice". The Philosophical Review 82 (2), 220–234.

Nozick, R. (1974): Anarchy, State, and Utopia, New York: Basic Books.

Rawls, J. (1951): "Outline of a Decision Procedure for Ethics". The Philosophical Review 60 (2), 177–197.

Rawls, J. (1971/99): A Theory of Justice, rev. ed., Oxford: Oxford University Press.

Rawls, J. (1974/75): "The Independence of Moral Theory". Proceedings and Addresses of the American Philosophical Association 48, 5–22.

Rawls, J. (2001): Justice as Fairness. A Restatement, ed. by Kelly, E., Cambridge (Mass.) / London: Harvard University Press.

Raz, J. (1982): "The Claims of Reflective Equilibrium". Inquiry 25 (3), 307–330.

Rescher, N. (1983): Risk. A Philosophical Introduction to the Theory of Risk Evaluation and Management, Washington: University Press of America.

Rutgers, B. (1998): "The Use of the Reflective Equilibrium Method in Normative Veterinary Ethics", in: van der Burg / van Willigenburg 1998a, 231–237.

Sandel, M.J. (1982): Liberalism and the Limits of Justice, Cambridge: Cambridge University Press.

Scanlon, T.M. (2003): "Rawls on Justification", in: Freeman, S. (ed.): The Cambridge Companion to Rawls, Cambridge: Cambridge University Press, 139–167.

Shackle, G.L.S. (1955): Uncertainty in Economics and Other Reflections, Cambridge: Cambridge University Press.

Simon, R.L. (1992): "Social and Political Philosophy", in: Becker, L.C. / Becker, C.B. (eds.): Encyclopedia of Ethics, Vol. 2, Chicago / London: St. James Press, 1163–1170.

Singer, P. (1974): "Sidgwick and Reflective Equilibrium", in: Singer, P.: Unsanctifying Human Life. Essays on Ethics, ed. by Kuhse, H., Oxford: Blackwell Publishers 2002, 27–50.

Solomon, W.D. (1995): "Ethics. III. Normative Ethical Theories", in: Post, S.G. (ed.): Encyclopedia of Bioethics, 3rd ed., Vol. 2, New York: Macmillan Reference USA 2004, 812–824.

Strong, C. (2010): "Theoretical and Practical Problems with Wide Reflective Equilibrium in Bioethics". Theoretical Medicine and Bioethics 31 (2), 123–140.

Tersman, F. (1993): Reflective Equilibrium. An Essay in Moral Epistemology, Stockholm: Almqvist & Wiksell International.

van den Beld, T. (1998): "Background Theories and Religious Beliefs: Their Role and Relation in Reflective Equilibrium", in: van der Burg / van Willigenburg 1998a, 73–88.

van der Burg, W. / van Willigenburg, T. (1998a) (eds.): Reflective Equilibrium. Essays in Honour of Robert Heeger, Dordrecht: Kluwer Academic Publishers.

van der Burg, W. / van Willigenburg, T. (1998b): "Introduction", in: van der Burg / van Willigenburg 1998a, 1–25.

van Thiel, G.J.M.W. / van Delden, J.J.M. (2010): "Reflective Equilibrium as a Normative Empirical Model". Ethical Perspectives 17 (2), 183–202.

van Willigenburg, T. (1998): "Morally Relevant Facts: Particularism and Intuitionist Rationality", in: van der Burg / van Willigenburg 1998a, 41–54.

Verweij, M. (1998): "Moral Principles: Authoritative Norms or Flexible Guidelines?", in: van der Burg / van Willigenburg 1998a, 29–40.

Widdershoven, G.A.M. (2007): "How to Combine Hermeneutics and Wide Reflective Equilibrium". Medicine, Health Care and Philosophy 10 (1), 49–52.