Theory and Decision (2023) 94:275–298
https://doi.org/10.1007/s11238-022-09889-6

Original position arguments and social choice under ignorance

Thijs De Coninck[1] · Frederik Van De Putte[2,3]

Accepted: 21 March 2022 / Published online: 30 May 2022

Abstract

John Rawls famously argued that the Difference Principle would be chosen by any rational agent in the original position. Derek Parfit and Philippe Van Parijs have claimed, contra Rawls, that it is not the Difference Principle which is implied by Rawls' original position argument, but rather the more refined Lexical Difference Principle. In this paper, we study both principles in the context of social choice under ignorance. First, we present a general format for evaluating original position arguments in this context. We argue that in this format, the Difference Principle can be specified in three conceptually distinct ways. We show that these three specifications give the same choice recommendations, and can be grounded in an original position argument in combination with the well-known maximin rule. Analogously, we argue that one can give at least four plausible specifications of the Lexical Difference Principle, which however turn out to give different recommendations in concrete choice scenarios. We prove that only one of these four specifications can be grounded in an original position argument. Moreover, this one specification seems the least appealing from the viewpoint of distributive justice. This insight points towards a general weakness of original position arguments.

Keywords Original position arguments · Difference principle · Leximin · Maximin · Social choice under ignorance

✉ Frederik Van De Putte
vandeputte@esphil.eur.nl; frederik.vandeputte@ugent.be

Thijs De Coninck
thijs.deconinck@ugent.be

[1] Centre for Logic and Philosophy of Science, Ghent University, Ghent, Belgium

[2] Erasmus Institute for Philosophy and Economics, Erasmus University of Rotterdam, Rotterdam, The Netherlands

[3] Centre for Logic and Philosophy of Science, Ghent University, Ghent, Belgium

1 Introduction

When a social planner chooses between different policies, there are two fundamental dimensions she needs to take into account:

(i) her uncertainty about the state of the world (including its underlying mechanisms) and, hence, about the consequences of her choices for society;
(ii) the welfare levels that will be enjoyed by the different members of society, relative to each specific state of the world and each policy.

For instance, when evaluating a national tax law that caps the higher incomes and redistributes the resulting financial resources, the social planner may be uncertain about the way the global economy will evolve—so she needs to consider the effects of such a law if the economy grows at a steady pace, but also if growth is hampered or worse. Moreover, she should consider, for each of these possibilities, how the tax law will affect the incomes of citizens, not only of the extremely rich, but also of various other classes in society.

Decision-making under uncertainty typically comes in two types: decision-making under ignorance and decision-making under risk. The latter refers to cases where we know the probabilities of each possible state (cf. Resnik 1987; Peterson 2017). In this paper, we limit ourselves to decision-making under ignorance. In particular, we study a number of specific rules for social choice under ignorance, that are all based on (variants of) John Rawls' famous Difference Principle (Rawls 1971). We ask when and how the rules in question give different choice recommendations, and if they can be grounded in a well-known type of argument, viz. original position arguments. In the remainder of this introduction, we clarify these terms and our overall aim.

Difference Principle(s) The Difference Principle states that we should "arrange social and economic inequalities in such a way that they are to the benefit of the least advantaged" (Rawls 1971, p. 266). According to Rawls, this principle should be applied to distributions of *primary goods*, which are "what persons need in their status as free and equal citizens, and as normal and fully cooperating members of society over a complete life" (Rawls 1971, p. xiii). Others, following Sen (1970) have understood the Difference Principle as specifying how we should choose between distributions of welfare or utility. In what follows, we interpret the principle in terms of welfare, though our insights apply *mutatis mutandis* to the interpretation in terms of primary goods as well.[1]

As an illustration of the Difference Principle, consider the basic scenario depicted in Fig. 1, where the social planner has to choose between three different alternatives *a*, *b*, and *c*. In this case, the Difference Principle recommends alternative *b* and

[1] Chung (2020) provides a formal analysis of Rawls' arguments against utilitarianism, showing that these arguments are ultimately self-defeating. A crucial point in Chung's analysis is Rawls' distinction between primary goods and the more abstract notions of utility and welfare (cf. Van Parijs (2001), pp. 10–11).

Fig. 1 Choice under certainty. Here, the couples (n, m) represent the welfare levels of the two individuals under consideration

Alternative	Welfare distribution
a	$(3, 1)$
b	$(2, 2)$
c	$(2, 3)$

c over alternative a because the least advantaged individual under b and c is better off than the least advantaged individual under a.[2]

According to Rawls, the Difference Principle is not only intuitively appealing, but it would also be applied by any rational decision-maker in the original position. The latter is understood as a situation in which the decision-maker is placed behind a "veil of ignorance", and so fully ignorant about her own position and the level of welfare she will enjoy under any alternative. In the context of our example in Fig. 1, it means that the decision-maker does not know whether under alternative a her personal welfare is given by 1 or 3. Rawls argues that given such a "fair" initial setup of the choice problem, and for the specific type of institutional choices that he focuses on, the decision-maker would choose the same alternatives as recommended by the Difference Principle. In this sense, the Difference Principle is said to be "grounded" in an original position argument.[3]

One critique of the Difference Principle, as noted by Sen (1970), is that it violates the strong Pareto Principle.[4] In our example alternative b is not Pareto optimal but recommended by the Difference Principle. The *Lexical Difference Principle* has been put forward as an alternative candidate which is consistent with the strong Pareto Principle. It states that one should first maximize the welfare of the worst-off individuals and, in case of equal welfare, maximize the welfare of the second worst-off individuals, and so on. On the Lexical Difference Principle, only alternative c would be recommended in our example. Parfit (1991) and Van Parijs (2001) claim that Rawls' original position argument supports the Lexical Difference Principle, rather than the Difference Principle.[5] Intuitively, this seems to make sense: the individual may turn out to be among the worst-off—in which case either principle

[2] Note that here, there is no uncertainty on the part of the social planner: every alternative is associated with a unique welfare distribution. We will show how the Difference Principle can be applied in cases of uncertainty in Sect. 3.

[3] See Gaus and Thrasher (2015) for an elaborate, informal analysis of original position arguments and their evolution in the work of Rawls and others. We give a formally precise characterization of what it means that a social choice rule is grounded in an original position argument in Sect. 2.4.

[4] On the strong Pareto Principle, a welfare distribution is optimal if and only if there is no other distribution such that no one is worse off and at least one individual is strictly better off under that other distribution.

[5] Parfit (1991, p. 38): "More importantly, this [the lexical difference principle] is the view to which we are led by Rawls's main arguments. From the standpoint of the original position, we would clearly favour giving benefits to the better off, when this would not worsen the position of those who are worse off. For all we know, we might be the people who are better off." Van Parijs (2001, p. 9): "According to this criterion [the lexical difference principle], inequalities are fine as long as they do not hurt the worst off or, if the worst off category is unaffected, the worst off category but one, etc. [...] this criterion is consistent with efficiency and fits easily into an original-position argument."

will maximize her welfare—but she may also turn out to be better off—in which case only the Lexical Difference Principle ensures that her welfare is maximized (conditional on maximizing the welfare of all the worse-off).

This paper We will study original position arguments for (Lexical) Difference Principles, in the context of social choice under ignorance. In doing so, we stick to Rawls' basic assumptions about the original position as much as possible. In particular, we assume interpersonal comparability of welfare levels on an ordinal scale, we exclude any information about the likelihood of states, and we exclude any probabilistic way of handling uncertainty as e.g. given by the well-known principle of insufficient reason.[6] In terms of the famous Rawls vs. Harsanyi-debate (Harsanyi 1975), this means that we side with Rawls on how to represent the reasoning and uncertainty of individuals in the original position.[7] We do so for the sake of the argument: it turns out that even if we grant all this, the prospects for grounding a Lexical Difference Principle in an original position argument are still fairly meagre. The upshot is that, while we agree with Parfit and Van Parijs that the Difference Principle should be strengthened in order to accommodate more "lexical" intuitions, it is not clear what such a strengthening looks like in the case of genuine ignorance about the state of affairs, and how such a strengthening can be grounded in an original position argument.

Our contribution can be summarized as follows. In Sect. 2, we present a general format for individual and social choice rules, we recall the well-known maximin and leximin rules for individual choice, and we show how original position arguments can be formalized. In Sect. 3, we argue that in the context of choice under ignorance, there are three approaches to social choice that are in line with the Difference Principle: a basic approach, an ex ante approach, and an ex post approach. After introducing these three approaches in general terms, we single out three corresponding social choice rules that all incorporate the Difference Principle in some sense. We show that, while they are conceptually distinct, these three social choice rules give the same choice recommendations and can be grounded in an original position argument in combination with the well-known maximin rule.

Analogously, in Sect. 4 we discuss specifications of the Lexical difference principle. Here, it turns out that the ex post approach can be further subdivided into two distinct approaches. Moreover, the four resulting social choice rules turn out to give distinct choice recommendations. As we explain, the specification of the Lexical difference principle that follows the basic approach can be grounded in an original position argument using leximin as the underlying individual choice rule. In contrast, none of the three more refined specifications can be grounded in *any* original position argument, regardless of the underlying notion of individual rationality. On the basis of these results and their proof, we conclude that original

[6] This principle tells us to treat each state as equally likely. Formally, where S is the set of possible states, we assign probability $\frac{1}{|S|}$ to each member of S. Once there, we can rank choices in terms of their expected utility (cf. Peterson 2017). Other probabilistic but more risk-averse ways to handle uncertainty in the original position can be found in Buchak (2017), Stefánsson (2019).

[7] See our discussion after Definition 7 (Sect. 2.4) for the similarities and differences between our formal account of original position arguments and that by Harsanyi.

position arguments—at least following our characterization—face important shortcomings in the context of choice under ignorance (Sect. 5).

Related work Most of the formal literature on Rawls is focused on the Rawls/Harsanyi debate over how exactly to characterize the original position and how agents would choose, once placed in such a situation (cf. supra) Moehler 2018; Moreno-Ternero and Roemer 2008; Roemer 2002. Within this debate, it is often presupposed that there is no uncertainty about the state of affairs, before the veil of ignorance is imposed.

Recall that Rawls claims that the Difference Principle would be chosen by any rational person in the original position. Rawls (1974) suggests that this claim could be supported by formal proof and considers the work of Arrow and Hurwicz (1972) to be a step in that direction. In Arrow and Hurwicz (1972), it is shown that if a decision rule satisfies certain plausible axioms it only takes into account the worst and best outcomes of each alternative to rank them. However, Arrow and Hurwicz do not really formalize the notion of an original position argument itself.

While there is a substantive literature on social choice under uncertainty, most of it focuses on cases of risk, i.e. cases where we know the probabilities of each possible state of affairs (cf. Mongin and Pivato 2021; Ben-Porath et al. 1997; Fleurbaey 2018; Gajdos and Maurin 2004; Gajdos and Kandil 2008; Hayashi and Lombardi 2019; Bovens 2015). As will become clear, our distinctions and examples bear many similarities with this strand of work. However, as is often the case, the devil is in the details.

We refer to Maskin (1979) for a general, axiomatic characterization of individual choice rules under ignorance. As he indicates, these axiomatizations are strongly linked to results in social choice theory, but Maskin does not consider the issue of social choice under ignorance per se, let alone the Rawlsian notion of an original position.

In Gustafsson (2018), Gustafsson argues that in situations of risk, a rational individual in the original position will not always make choices that agree with the Difference Principle—whether it is spelled out according to an ex ante or an ex post approach. In this sense, although his argument concerns decisions under risk and ours concerns decisions under ignorance, we agree with Gustafsson that there is a mismatch between the original position on the one hand, and principles of justice on the other. However, while Gustafsson uses this insight against the Difference Principle, our conclusion is in a sense the opposite: we argue that, since the most natural model of individual reasoning in the original position cannot accommodate certain distinctions that seem to be crucial for our intuitive notion of justice, this model should be refined if not revised. We leave the latter enterprise for future work.

In Strasnick (1976), Strasnick argues that the Difference Principle follows naturally from the properties of the original position once suitably defined. While our modelling choices differ from his, our work can be conceived as a continuation of this line of research. A key difference with our work is that Strasnick assumes social choice in the absence of any uncertainty. As will become clear, it is precisely the dimension of uncertainty that causes trouble for arguments based on the original position.

2 Choice scenarios and the original position

In this section, we introduce the models that we will be working with throughout this article, viz. choice scenarios (Sect. 2.1). We discuss the maximin and leximin decision rule in terms of these models (Sect. 2.2) and introduce some general properties of individual and social choice rules (Sect. 2.3). Finally, we define a transformation that turns every choice scenario into one that has the characteristics of the original position (Sect. 2.4).

2.1 Choice scenarios

The models that we work with are obtained by combining well-known ingredients from the study of welfare distributions (cf. Sen 1970) and from decision-making under ignorance (cf. Resnik 1987; Peterson 2017).

Definition 1 A *choice scenario* is a tuple $\mathfrak{C} = \langle N, A, S, d \rangle$, where N is a non-empty finite set of *individuals*, A a non-empty finite set of *alternatives*, S a non-empty finite set of *states*, and $d : N \times A \times S \rightarrow \mathbb{R}$ a *welfare distribution function*.

In a given choice scenario, the set S represents the uncertainty of the decision-maker, i.e. S is the set of states the decision-maker considers possible. If S is a singleton we call the choice scenario *basic*. The members of $A \times S$ are called the (possible) *outcomes* of the scenario. For each of these outcomes (a, s) and each individual $i \in N$, the distribution function d determines the welfare of each individual $i \in N$.[8] To simplify notation, we write $d_i(a, s)$ instead of $d(i, a, s)$ and use $\overline{d(a, s)}$ to denote the $|N|$-tuple $\langle d_i(a, s) \rangle_{i \in N}$.[9] A scenario in which, for all outcomes (a, s), $d_i(a, s) = d_j(a, s)$ for all $i, j \in N$, will be called *egalitarian*. A scenario in which N is a singleton will be called an *individual choice scenario*. Hence, individual choice scenarios are by definition also egalitarian.

Note that our models assume that we can compare individuals on the basis of these welfare levels. In other words, we take it that, depending on the context, the welfare levels represent something commensurable, such as Rawls's bundles of primary goods, monetary status, etc.

Figure 2 represents a simple choice scenario with two individuals i and j, three alternatives and two states. Here, the couples (n, m) represent the distribution function, where $n = d_i(a, s)$ and $m = d_j(a, s)$. As a convention, we always list the welfare of i before that of j in our figures. For example, at outcome (b, s_1) individual i's welfare is 1 whereas individual j's welfare is 2 so that individual j is considered

[8] We use real numbers to represent welfare levels, following common practice in decision theory. However, as all our models are finite, nothing hinges on this and one might as well use natural numbers.

[9] When using this type of notation, we assume fixed an ordering of N (further below we use similar notation assuming that also S is an ordered set). This assumption is harmless, given that N (S) is finite.

Fig. 2 The choice scenario $\mathfrak{C}_1$ with two states and three alternatives

$\mathfrak{C}_1$	s_1	s_2
a	$(1,1)$	$(2,2)$
b	$(1,2)$	$(1,3)$
c	$(3,0)$	$(1,1)$

better off than individual i at that outcome. We will use this choice scenario as our running example in Sects. 2 and 3.

2.2 Choice rules

Given some choice scenario, a *choice rule* is used to select a subset of the alternatives that are considered to be *admissible* in some specific sense. We distinguish between *social choice rules* and *individual choice rules*. By the latter, we mean any choice rule that specifies admissibility for one individual $i \in N$ as a function (solely) of the welfare of i at the various outcomes in the scenario.

In line with our focus on the Difference Principle and its lexical variations, we focus on two individual choice rules and the corresponding notions of admissibility: the *maximin* rule and the *leximin* rule.[10] To introduce the former, we first fix some notation.

Notation 1 Where X is a finite subset of $\mathbb{R}$, we write $\mathsf{min}(X)$ to denote the $\leq$-minimal element of X, i.e. $x = \mathsf{min}(X)$ iff $x \in X$ and for all $y \in X : y \geq x$.

Definition 2 (*Maximin ranking*) Let $\mathfrak{C} = \langle N, A, S, d \rangle$ be a choice scenario, $i \in N$, and $a, b \in A$:

$$a \succeq_i^{\mathsf{max}} b \quad \text{iff} \quad \mathsf{min}\{d_i(a,s) \mid s \in S\} \geq \mathsf{min}\{d_i(b,s) \mid s \in S\}$$

We use $\succ_i^{\mathsf{max}}$ to denote the strict counterpart of $\succeq_i^{\mathsf{max}}$.

Definition 3 (*Maximin admissibility*) Where $\mathfrak{C} = \langle N, A, S, d \rangle$ is a choice scenario, $i \in N$, and $a \in A$, a is *maximin admissible for i* if and only if for all $b \in A : a \succeq_i^{\mathsf{max}} b$.

For example, in the scenario depicted in Fig. 2, alternative a is not maximin admissible for individual j because the minimum welfare that j can receive given b is 2 while the minimum welfare given alternative a is 1. Likewise, alternative c is not maximin admissible for j. In contrast, all alternatives are maximin admissible for i.

Maximin is a relatively weak rule. In our running example, alternatives b and c are equally preferred for i under $\succeq_i^{\mathsf{max}}$, since the worst welfare i can receive under

[10] See (Peterson 2017, Chapter 3) and (Resnik 1987, Chapter 2) for a survey of other well-known decision rules under ignorance, including the optimism–pessimism rule, weak dominance, and minimax regret. It is an open question whether and how our findings in this paper transfer to those rules. See also Maskin (1979) for an axiomatic characterization of i.a. the maximin and leximin rule within the context of choice under ignorance.

both is 1. This is not very intuitive since, if we were to look at the second worst outcome for individual i, then alternative c is clearly better than alternative b. *Leximin* is often presented as a refinement of maximin that takes this intuition into account. Informally, it says that if the worst outcomes of two alternatives are equal, one should choose the alternative such that the second worst outcome is better than the second worst outcome of the other alternative. If the second worst outcomes are equal then one should look at the third worst outcomes, and so on, until either one is better than the other or they are completely equal. To define leximin in exact terms, it will be useful to introduce some extra notation.

Notation 2 Let $\overline{x} \in \mathbb{R}^m$. We write $\overrightarrow{x}$ for the m-tuple that contains all the items in $\overline{x}$ ordered from smallest to largest.[11] Where $1 \leq n \leq m$, let $\overline{x}(n)$ denote the n^{th} entry of $\overline{x}$.

Where $\mathfrak{C} = \langle N, A, S, d \rangle$ is a choice scenario, $i \in N$, and $a \in A$, let

$$\overline{d_i(a)} =_{\mathsf{df}} \langle d_i(a,s) \rangle_{s \in S}$$

be the tuple including all welfare values individual i can receive given alternative a. In line with the above, $\overrightarrow{d_i(a)}$ is the ordered counterpart of $\overline{d_i(a)}$. Before we define the leximin ranking on alternatives, it is useful to define the lexicographic ranking for arbitrary tuples of real numbers.

Definition 4 (*Lexicographic ranking*) Where $\overline{x}, \overline{y} \in \mathbb{R}^m$: $\overline{x} \sqsupseteq_{\mathsf{lex}} \overline{y}$ if and only if either $\overline{x} = \overline{y}$ or there is an $n \leq m$ such that (i) for all $k < n$, $\overline{x}(k) = \overline{y}(k)$ and (ii) $\overline{x}(n) > \overline{y}(n)$. We call $\sqsupseteq_{\mathsf{lex}}$ the *lexicographic ranking*. Its strict counterpart is denoted by $\sqsupset_{\mathsf{lex}}$.

Definition 5 (*Leximin ranking of alternatives*) Where $\mathfrak{C} = \langle N, A, S, d \rangle$ is a choice scenario, $i \in N$, and $a, b \in A$:

$$a \succeq_i^{\mathsf{lex}} b \quad \text{iff} \quad \overrightarrow{d_i(a)} \sqsupseteq_{\mathsf{lex}} \overrightarrow{d_i(b)}$$

We use $\succ_i^{\mathsf{lex}}$ to denote the strict counterpart of $\succeq_i^{\mathsf{lex}}$.

Definition 6 (Leximin admissibility) Where $\mathfrak{C} = \langle N, A, S, d \rangle$ is a choice scenario, $i \in N$, and $a \in A$, a is *leximin admissible for i* if and only if for all $b \in A : a \succeq_i^{\mathsf{lex}} b$.

In our running example, we have $c \succ_i^{\mathsf{lex}} a \succ_i^{\mathsf{lex}} b$, since $\overrightarrow{d_i(c)} = \langle 1, 3 \rangle$, $\overrightarrow{d_i(a)} = \langle 1, 2 \rangle$, $\overrightarrow{d_i(b)} = \langle 1, 1 \rangle$ and $\langle 1, 3 \rangle \sqsupset_{\mathsf{lex}} \langle 1, 2 \rangle \sqsupset_{\mathsf{lex}} \langle 1, 1 \rangle$. Hence, only c is admissible for i. It can easily be verified that, in general, $\succ_i^{\mathsf{max}} \subseteq \succ_i^{\mathsf{lex}}$, and often the inclusion is proper. As a result, any leximin admissible alternative for i is also maximin admissible for i, while the converse often fails (cf. our running example). In this sense, leximin is a refinement of maximin.

[11] If some $y \in \mathbb{R}$ occurs several times in $\overline{x}$, then y also occurs as many times in $\overrightarrow{x}$. So for instance, where $\overline{x} = \langle 1, 2, 3, 2 \rangle$, we have $\overrightarrow{x} = \langle 1, 2, 2, 3 \rangle$.

$\mathfrak{C}_1$	s_1	s_2
a	$(1,1)$	$(2,2)$
b	$(1,2)$	$(1,3)$
c	$(3,0)$	$(1,1)$

$\mathfrak{C}_2$	t_1	t_2
a	$(1,1)$	$(2,2)$
b	$(1,2)$	$(1,3)$
c	$(3,0)$	$(1,1)$

$\mathfrak{C}_3$	t_1	t_2
d	$(1,2)$	$(1,3)$
e	$(1,1)$	$(2,2)$
f	$(3,0)$	$(1,1)$

Fig. 3 Scenario $\mathfrak{C}_2$ is obtained from $\mathfrak{C}_1$ by relabelling the states and scenario $\mathfrak{C}_3$ can be obtained from $\mathfrak{C}_2$ by relabelling the alternatives

2.3 Conditions on choice rules

In what follows we discuss three conditions on choice rules: individualism, column symmetry, and row symmetry. These conditions will play an important role later on (cf. Sect. 4.2).

Individualism requires that given some individual any changes to the welfare of other individuals does not affect which alternatives are admissible for that individual. To be more precise, given some choice scenario $\mathfrak{C} = \langle N, A, S, d \rangle$ and individual $i \in N$, if we change the payoffs of any of the other individuals $j \neq i$, then individualism requires that nothing changes to the admissibility of alternatives for i.

Column symmetry states that a choice rule should not be sensitive to the way states are labeled but only to the payoffs individuals receive in those states. Finally, row symmetry says that the labeling of alternatives is irrelevant. The latter two principles taken together are known as Milnor (1954) symmetry condition or Arrow and Hurwicz's property B (Arrow and Hurwicz 1972). Both column and row symmetry are considered standard when dealing with choice under ignorance, i.e. when there are no probabilities associated with states of nature (cf. Peterson 2017). We illustrate these symmetry conditions with an example, cf. Fig. 3.

By column symmetry a is admissible in $\mathfrak{C}_1$ iff a is admissible in $\mathfrak{C}_2$ since the only thing differentiating $\mathfrak{C}_1$ from $\mathfrak{C}_2$ are the labels of states, i.e. s_1 corresponds to t_1 and s_2 to t_2. By row symmetry a is admissible in $\mathfrak{C}_2$ iff e is admissible in $\mathfrak{C}_3$ since $\mathfrak{C}_3$ can be obtained from $\mathfrak{C}_2$ by relabelling the alternatives. Putting it in more technical terms, whenever we have a bijection $\sigma : A \rightarrow A'$ (or $\sigma : S \rightarrow S'$ for column symmetry) that preserves payoffs[12] for all $a \in A$ then by row symmetry a is admissible iff $\sigma(a)$ is admissible.

In the remainder of this article, we assume that individual choice rules satisfy individualism and column symmetry (we rely on these properties in Sect. 4.2). While row symmetry is not required for our results to go through, all particular choice rules that will be defined satisfy this condition as well.

2.4 The original position

In the original position, individuals make their choices behind a veil of ignorance: they do not know what position they will occupy in society, once an alternative has been chosen. Moreover, the individuals do not even have any probabilistic

[12] I.e. for all $a \in A$, $i \in N$ and $s \in S$: $d(i, a, s) = d'(i, \sigma(a), s)$.

information about their possible positions (Rawls 1971, p. 134). In other words, the individuals are fully ignorant about their own position, and hence about the level of welfare they will enjoy.[13] Given this "fair" initial setup of the choice problem, the thought is that whatever principle of justice is chosen by any of the individuals, it will be fair as well. This is essentially how Rawls argues for his principles of justice in Rawls (1971).

Recall that our overall aim is to inspect whether particular social choice rules can be grounded in a similar original position argument. To do so, we start from the view that there is not a single choice scenario that corresponds to the original position; instead, for each particular choice scenario, we can construct a corresponding choice scenario which has the characteristics of an original position. We call the latter the *original position transformation* (abbreviation: OP-transformation) of the original choice scenario. It is defined as follows:

Definition 7 (OP-transformation) Let $\mathfrak{C} = \langle N, A, S, d \rangle$ be a choice scenario. Let Π be the set of all bijective functions $\pi : N \to N$. The *original position transformation* of $\mathfrak{C}$ is the choice scenario $\mathfrak{C}^* = \langle N, A, S^*, d^* \rangle$, where

- $S^* = S \times \Pi$
- for all $i \in N$, $a \in A$, and $(s, \pi) \in S^*$: $d^*(i, a, (s, \pi)) = d(\pi(i), a, s)$

In other words, given some choice scenario $\mathfrak{C}$, we obtain its OP-transformation $\mathfrak{C}^*$ by combining the ignorance in the original model with ignorance about the individual's "identities", and the way these identities affect the level of welfare one gets. That $\pi(i) = j$ at a certain state (s, π) in $\mathfrak{C}^*$ can be interpreted, loosely speaking, as saying that i occupies the position that is originally assigned to j in s, and hence enjoys j's welfare at this state. One might as well introduce a separate category of positions, assign individuals to positions, and take the distribution of welfare to depend on the positions, but this would not affect any of our results and only complicate notation and terminology.[14]

Importantly, the OP transformation of any choice scenario is just another choice scenario. Thus, in particular, we can apply individual choice rules to the OP transformation of any given choice scenario. This is crucial, if we want to ground social choice rules using an original position argument in combination with an individual choice rule.

Let us illustrate the above by means of our running example. Figure 4 displays, on the left-hand side, the choice scenario $\mathfrak{C}_1$, and on the right hand side, its OP transformation. Here, $\pi_=$ is simply the identity relation, and $\pi_{\neq}$ swaps the two agents, i.e. $\pi_=(i) = i$, $\pi_=(j) = j$, $\pi_{\neq}(i) = j$, and $\pi_{\neq}(j) = i$. If we apply the maximin rule to the OP transformation, we find that both a and b are admissible for i, while c is not.

[13] See Gajdos and Kandil (2008) for a different approach to ignorance in the Original Position, where complete ignorance is modelled as considering all probability distributions over identities and states possible.

[14] See however De Coninck and Van De Putte (2021) for a formal logical investigation into original position arguments, which does take such positions as primitive.

$\mathfrak{C}_1$	s_1	s_2
a	$(1,1)$	$(2,2)$
b	$(1,2)$	$(1,3)$
c	$(3,0)$	$(1,1)$

$\mathfrak{C}_1^*$	$(s_1,\pi_=)$	$(s_2,\pi_=)$	$(s_1,\pi_{\neq})$	$(s_2,\pi_{\neq})$
a	$(1,1)$	$(2,2)$	$(1,1)$	$(2,2)$
b	$(1,2)$	$(1,3)$	$(2,1)$	$(3,1)$
c	$(3,0)$	$(1,1)$	$(0,3)$	$(1,1)$

Fig. 4 A choice scenario ($\mathfrak{C}_1$) and its OP-transformation ($\mathfrak{C}_1^*$)

As we can see, in $\mathfrak{C}_1^*$ both individuals i and j have the same ordered list of welfare values, i.e. $\overrightarrow{d_i^*(\alpha)} = \overrightarrow{d_j^*(\alpha)}$ for $\alpha \in \{a,b,c\}$. This holds more generally:

Observation 1 Let $\mathfrak{C}^* = \langle N, A, S^*, d^* \rangle$ be the original position transformation of a choice scenario, let $a \in A$ and let $i,j \in N$. Then $\overrightarrow{d_i^*(a)} = \overrightarrow{d_j^*(a)}$.

By Observation 1 and relying on column symmetry and individualism (cf. Sect. 2.3), one can easily show that for any OP-transformation, admissibility on any individual choice rule is uniform across all individuals. In our example, this means that also for j, only a and b are admissible in the OP transformation of $\mathfrak{C}_1$. More generally:[15]

Observation 2 Let $\mathfrak{C}^* = \langle N, A, S, d \rangle$ be the OP-transformation of a choice scenario, let $a \in A$ and let $i,j \in N$. Let $\mathbf{R}$ be an individual choice rule. Then a is $\mathbf{R}$-admissible for i in $\mathfrak{C}^*$ if and only if a is $\mathbf{R}$-admissible for j in $\mathfrak{C}^*$.

With the above terminology in hand, we can specify a social choice rule $\mathbf{S}^{\mathrm{O}}(\mathbf{R})$ starting from any individual choice rule $\mathbf{R}$. That is, say an option a is socially admissible in $\mathfrak{C}$ according to $\mathbf{S}^{\mathrm{O}}(\mathbf{R})$, iff a is $\mathbf{R}$-admissible for any (and hence every) individual $i \in N$ in $\mathfrak{C}^*$. For example, on the maximin rule, alternatives a and b are admissible in scenario $\mathfrak{C}_1^*$ and hence we may say that a and b are socially admissible in $\mathfrak{C}_1$ according to $\mathbf{S}^{\mathrm{O}}(\mathit{maximin})$.

So any individual choice rule corresponds to a social choice rule, using the OP transformation. In what follows, we will focus on correspondence in the other direction. That is, given a social choice rule $\mathbf{S}$, we will ask whether there is an individual choice rule $\mathbf{R}$ such that $\mathbf{S} = \mathbf{S}^{\mathrm{O}}(\mathbf{R})$. In that case, we say that we can *ground* the social choice rule in an original position argument. In particular, we will investigate social choice rules that are motivated by (variants of) the Difference Principle, and ask whether and how they can be shown to agree with $\mathbf{S}^{\mathrm{O}}(\mathit{maximin})$, with $\mathbf{S}^{\mathrm{O}}(\mathit{leximin})$, or with $\mathbf{S}^{\mathrm{O}}(\mathbf{R})$ for any other individual choice rule $\mathbf{R}$.

To finish this section, let us briefly compare our account of the original position with that of Harsanyi (1953, 1977). Like Harsanyi's impartial observer model, we assume that introducing a veil of ignorance involves expanding the state space, by

[15] Observation 2 concurs with the analysis of original position arguments by Gaus and Thrasher (Gaus and Thrasher 2015, pp. 39–43), who argue that there is no need for consent, bargaining, or aggregation across individuals, once the decision-makers are placed behind the veil of ignorance. They write: 'once difference has been eliminated, the justification in the original position is via an individual principle of rational parametric choice.' (Gaus and Thrasher 2015, p. 42).

taking its product with the possible set of permutations of individuals. Unlike Harsanyi, we do not assume any probabilistic expectations about the relative likelihood that people end up in certain positions.

3 The difference principle

In this section, we argue that a number of conceptually distinct specifications can be given of Rawls' Difference Principle, when conceived as a social choice rule and in the context of choice under ignorance (Sect. 3.1). Some of these specifications turn out to be equivalent, others are not. We then single out those specifications that are validated by the original position argument where the individuals in question apply the maximin rule (Sect. 3.2).

3.1 The difference principle: three approaches

If choices are deterministic, i.e. in the absence of uncertainty, it is pretty clear what the Difference Principle recommends us to do (cf. Sect. 1). In contrast, in a situation where the state of the world is uncertain, there are several ways one can spell out the Difference Principle in exact terms. For choice under risk, i.e. where a probability distribution over S is given, one can e.g. distinguish between maximizing the minimal ex ante welfare, maximizing the ex post minimal welfare, or using a "mixed rule" (Mongin and Pivato 2021).

In what follows, we show that analogous distinctions can be made for choice under ignorance. We distinguish in particular between three general approaches, which we call the basic approach, the ex ante approach, and the ex post approach.[16] We will first explain the general idea behind these approaches in informal terms, after which we show how they can be instantiated by concrete social choice rules. This way we also set the stage for our discussion of lexical social choice rules in Sect. 4.

Basic approach When deciding what the value of a given alternative $a \in A$ is, we could simply ignore the distinction between different states and different individuals, and treat all welfare levels $d_i(a, s)$ (for some $s \in S$ and some $i \in N$) as interchangeable. This means that the value of a is a function of the tuple

[16] The main difference with the distinctions in Mongin and Pivato (2021) can be understood as follows: while Mongin and Pivato only consider one "pure" ex post approach, we distinguish between two: one where we simply focus on the ex post worst off agent at each state, and one in which we look at the ex post distributions, evaluate those (in terms of a social welfare index), and then aim for alternatives that maximally promote the social welfare. While the difference is merely conceptual in this section, we do arrive at distinct rules as soon as we apply these approaches in a lexical fashion, as we show in Sect. 4.

$$\overline{d_{\mathsf{b}}(a)} =_{\mathsf{df}} \langle d_i(a,s)\rangle_{i\in N, s\in S}$$

For instance, on the maximin version of this approach, the value of a is given by the smallest $x \in \mathbb{R}$ that occurs in $\overline{d_{\mathsf{b}}(a)}$. On the leximin version, the value[17] of a is given by the ordered tuple $\overrightarrow{d_{\mathsf{b}}(a)}$. For our running example (Fig. 2), we see that on the maximin approach the value of a and b is 1, while the value of c is 0. Hence, only a and b are difference admissible on the maximin version of this approach. On the leximin approach, we have $\overrightarrow{d_{\mathsf{b}}(a)} = \langle 1,1,2,2\rangle$, $\overrightarrow{d_{\mathsf{b}}(b)} = \langle 1,1,2,3\rangle$, and $\overrightarrow{d_{\mathsf{b}}(c)} = \langle 0,1,1,3\rangle$. Hence, only alternative b is admissible on the leximin version of this approach.

As should be clear to the reader, any combination of the basic approach with an individual choice rule $\mathbf{R}$ will give us a social choice rule $\mathbf{S}^{\mathsf{b}}(\mathbf{R})$ that is equivalent to the social choice rule obtained by combining $\mathbf{R}$ with the OP-transformation—i.e., $\mathbf{S}^{\mathsf{b}}(\mathbf{R}) = \mathbf{S}^{\mathsf{o}}(\mathbf{R})$.[18] In this sense the basic approach can be seen as a mere reformulation of the original position argument.

Ex ante approach According to the ex ante interpretation of "the worst-off", we first need to determine how well-off each individual is, relative to each alternative, and given our ignorance about the actual state of affairs. Let us call this the *expected welfare* of i given a. Mind that, since we are working in the context of ignorance, this informal notion is not to be confused with expected utility in the standard, technical sense. We just use the notion here as a placeholder for any more specific way one may assign some index of welfare to a given alternative $a \in A$, in view of the welfare levels that i receives at each state s given a.

Individual i can then be said to be worst-off given a if and only if the expected welfare of i is worst among all individuals (again, given a). The *value* of a given alternative a is then a function of the expected welfare of the worst-off individual given a, and used to compare a with other alternatives.

As we already noted, the notion "expected welfare" can be specified in many ways. Here, we focus on two examples and use them to illustrate the overall idea behind the ex ante approach:

- *Pessimistic* expected welfare: define the expected welfare of i given a as $\mathsf{min}\{d_i(a,s) \mid s \in S\}$. Next, simply compare the expected welfare of different individuals (given an alternative a) using the standard "greater than" order on the reals. Finally, define the value of a as the expected welfare of the worst-off individual, and use that value to determine which alternatives are socially admissible. Using our running example, the pessimistic expected welfare of i under alternative b is 1, whereas for individual j it is 2. So the worst-off individual given b is i and the value of b is given by welfare level 1. Similarly,

[17] By the "value" of an alternative with respect to some choice rule we mean the information that is relevant to decide on the ranking of that alternative. We do not presuppose that this value is representable by a real number.

[18] Mind that for this to hold, we need to rely again on the conditions for individual choice rules, cf. Sect. 2.2.

Fig. 5 The matrix on the left depicts our running example while the matrix on the right is obtained by considering the welfare of the abstract individual w

$\mathfrak{C}_1$	s_1	s_2
a	$(1,1)$	$(2,2)$
b	$(1,2)$	$(1,3)$
c	$(3,0)$	$(1,1)$

$\mathfrak{C}_1^{\mathsf{w}}$	s_1	s_2
a	1	2
b	1	1
c	0	1

the value of a is given by 1 and the value of c by 0. Consequently, only alternatives a and b are socially admissible on this account.

- *Lexical* expected welfare: define the expected welfare of i given a as the ordered tuple $\overrightarrow{d_i(a)}$. Next, compare the expected welfare of different individuals using the lexicographic ordering: i's expected welfare given a is at least as good as j's expected welfare given a if and only if $\overrightarrow{d_i(a)} \sqsupseteq_{\mathsf{lex}} \overrightarrow{d_j(a)}$. Once there, define the value of a as the expected welfare of the worst-off individual, i.e. the lowest-ranked individual on the lexicographic order. This will be a tuple of welfare levels. We then compare alternatives in terms of such tuples, by applying the lexical ranking all over again. Applying this to the running example, the expected welfare of alternative b for i is $\langle 1, 1\rangle$ and for j it is $\langle 2, 3\rangle$. Hence individual i is worst-off given alternative b. Some calculation yields that only a is socially admissible on this account, since $\langle 1, 2\rangle \sqsupset_{\mathsf{lex}} \langle 1, 1\rangle \sqsupset_{\mathsf{lex}} \langle 0, 1\rangle$.

From the above, it is clear that the ex ante approach can be spelled out in various ways, some of which are logically distinct. In line with our observations from Sect. 2, the lexicographic interpretation of "ex ante worst-off" is stronger than the pessimistic one: if an individual is ex ante worst-off under the lexicographic interpretation, then she will also be worst-off on the pessimistic reading. As a result, these interpretations may also give distinct recommendations for certain choice scenarios.

Ex post approach Rather than taking the ex ante perspective, one can also interpret the notion "worst-off" in an ex post sense. Given a pair (a, s), it is clear who is (among the) worst-off. So instead of looking at all individuals, the difference principle would tell us to only consider the welfare of the (abstract) "individual" w, whose welfare at outcome (a, s) is given by $d_{\mathsf{w}}(a, s) = \mathsf{min}\{d_i(a, s) \mid i \in N\}$. This transforms the choice scenario into an individual one, as illustrated by Fig. 5.

Once the choice scenario has been transformed in this way, we can apply any individual choice rule. In our running example, both a and b are maximin admissible in the transformed scenario, while only a is leximin admissible.[19]

[19] In light of this example the reader may conjecture that, while conceptually distinct, the ex ante approach and the ex post approach always give the same recommendations. While this conjecture is true if we use maximin as the underlying basis for comparison (cf. Observation 3), it fails if we use leximin (this follows immediately from observations that we make in the last two paragraphs of Sect. 4.2).

3.2 Maximin in the original position

As we just argued, there is a range of conceptually distinct ways to specify the Difference Principle, once we allow for uncertainty. In what follows, we single out those specifications that are validated by an original position argument, when combined with the maximin rule. Doing so, we show that at least some versions of the Difference Principle can be grounded in this way.

The three specifications in question are: (i) the basic approach, in combination with maximin; (ii) the pessimistic ex ante approach; (iii) the ex post approach, combined with maximin. As for (i) this was already noted in Sect. 3.1. But also for (ii) and (iii), a little reflection shows that the value of an alternative ultimately boils down to the lowest welfare level $x \in \mathbb{R}$ such that, for some $i \in N$ and some $s \in S$, $x = d_i(a, s)$. In terms of decision matrices, all that matters for these specifications of the Difference Principle is the lowest number that occurs in some tuple in the row corresponding to the choice in question. This explains at once why these specifications of the Difference Principle are validated by the original position argument that takes maximin as a standard for individual rationality:

Observation 3 Let $\mathfrak{C} = \langle N, A, S, d \rangle$ be a social choice scenario and let $a \in A$. Then each of the following are equivalent:

- a is socially admissible in $\mathfrak{C}$ on the basic approach combined with maximin,
- a is socially admissible in $\mathfrak{C}$ on the ex ante approach in terms of pessimistic expected welfare,
- a is socially admissible in $\mathfrak{C}$ on the ex post approach combined with maximin,
- a is maximin admissible for every $i \in N$ in $\mathfrak{C}^*$.

Observation 3 should not come as a surprise. The interest lies here not so much in this observation as a mathematical result, but rather in its formal structure. In what follows, we will ask whether a structurally similar result can be proven for the Lexical Difference Principle, once it has been suitably specified. While this obviously holds for the basic approach, the interesting question is whether it also holds for ex ante and ex post versions.

4 The lexical difference principle

As noted in the introduction, Parfit (1991) and Van Parijs (2001) claim that a rational individual in the original position would prefer a lexical variant of the Difference Principle (cf. Sect. 1). When there is no ignorance (i.e. when S is a singleton), this is plausible: the social choice rule that derives from leximin in the original position maximizes *everyone's* welfare (conditional on maximizing the welfare of all those that are worse off), not just that of the worst-off. But how does this claim fare in general, i.e. for social choice under ignorance?

In Sect. 4.1, we specify and illustrate three social choice rules that can be said to incorporate the Lexical Difference Principle and are conceptually distinct from the basic approach (cf. Sect. 3.1). Using the running example in Fig. 6, we show that

Fig. 6 $\mathfrak{C}_2$, our running example for Sect. 4.1, featuring two individuals i, j, four alternatives a, b, c, d, and two states s_1, s_2

$\mathfrak{C}_2$	s_1	s_2
a	$(2,1)$	$(3,4)$
b	$(3,3)$	$(1,3)$
c	$(5,1)$	$(3,2)$
d	$(4,4)$	$(1,2)$

the four rules in question give distinct recommendations in cases of genuine ignorance. Next, we prove that none of these rules, except for the basic approach, can be grounded in any original position argument (Sect. 4.2). The same argument applies *mutatis mutandis* to the two "lexical" variants in Sect. 3. In sum, save for $\mathbf{S}^{\mathsf{b}}(\textit{leximin})$, no plausible lexical social choice rule can be grounded in an original position argument.

4.1 Lexical social choice rules

Our classification follows that of Sect. 3.1, with two differences. On the one hand, we do not redefine the basic approach, having already noted how that can be combined with any individual choice rule, including leximin. On the other hand, as we explain below, the ex post approach can be further subdivided into two categories that turn out to be conceptually and logically distinct.

Importantly, we specify the social choice rules in such a way that two basic desiderata are satisfied. First, when applied to a strictly egalitarian scenario, the rules give the same recommendation as the leximin rule would give to any of the individuals. Second, when applied to a basic scenario, they give the same recommendation as the leximin rule would give, when applied to the original position transformation of that scenario. Both desiderata seem natural if we wish to see these rules as incorporating the Lexical Difference Principle. They are moreover necessary, if we want the rules to be equivalent to leximin in the original position, at least for the subclass of scenarios that are strictly egalitarian, resp. basic.

The ex ante approach Recall that in this approach, we need to specify and compare the ex ante welfare of each individual given a certain alternative a. Here we take the lexicographic interpretation of the latter notion, using the ordered tuple $\overrightarrow{d_i(a)}$ to represent the ex ante welfare of i given a. We can then say that i is *worst-off* given a if and only if, for no $j \in N$, $\overrightarrow{d_i(a)} \sqsupset_{\mathsf{lex}} \overrightarrow{d_j(a)}$. In order to define the notions of (ex ante) second worst-off, third worst-off, etc. we first need some more notation.

Notation 3 Let $\overline{X} \in (\mathbb{R}^m)^n$ (i.e. $\overline{X}$ is an n-tuple of m-tuples in $\mathbb{R}$). We write $\overline{X}(k)$ to denote the k^{th} entry of $\overline{X}$. We let $\overrightarrow{X}$ denote the n-tuple obtained by ordering the tuples within $\overline{X}$ from "worst" to "best", according to the standard lexicographic order (i.e. $\sqsupseteq_{\mathsf{lex}}$).[20]

For every alternative $a \in A$, let $\overline{D_{\mathsf{e}}(a)} =_{\mathsf{df}} \langle \overrightarrow{d_i(a)} \rangle_{i \in N}$. Note that $\overline{D_{\mathsf{e}}(a)}$ is a tuple of ordered tuples of real numbers; more precisely, it is a member of the set $(\mathbb{R}^{|S|})^{|N|}$. If we order these tuples according to $\sqsupseteq_{\mathsf{lex}}$, we find out which individual has the worst expectations given a, who has the second worst expectations given a, etc. We may say that $i \in N$ is *k-worst-off* given a if and only if $\overrightarrow{d_i(a)} = \overrightarrow{D_{\mathsf{e}}(a)}(k)$. Note that several individuals may be k-worst-off given a certain alternative a—this will be so if there are several individuals whose ex ante welfare is the same and identical to the kth entry of $\overrightarrow{D_{\mathsf{e}}(a)}$.

To illustrate the above, let us have a look at the choice scenario in Fig. 6. Given a, we see that the individual with the worst expectations is j, since $\langle 2,3 \rangle \sqsupset_{\mathsf{lex}} \langle 1,4 \rangle$. Given b, individual i has the worst expectations, since $\langle 3,3 \rangle \sqsupset_{\mathsf{lex}} \langle 1,3 \rangle$. For c, it is again j who has the worst expectations, viz. $\langle 1,2 \rangle$, and for d it is again i.

Once there, recall that the overall aim is to make a choice that first maximizes the ex ante welfare of the worst-off, but in case of ties, we also look at the second worst-off, the third worst-off, and so on. So, we need to compare the tuples $\overrightarrow{D_{\mathsf{e}}(a)}, \overrightarrow{D_{\mathsf{e}}(b)}, \ldots$ lexicographically. The following definition shows how such tuples of tuples can be compared in general:[21]

Definition 8 Where $\overline{X}, \overline{Y} \in (\mathbb{R}^m)^n$, $\overline{X} \sqsupseteq^{\mathsf{m}}_{\mathsf{lex}} \overline{Y}$ if and only if either $\overline{X} = \overline{Y}$ or there is a $k \leq n$ such that (i) for all $l < k$, $\overline{X}(l) = \overline{Y}(l)$ and (ii) $\overline{X}(k) \sqsupset_{\mathsf{lex}} \overline{Y}(k)$. We call $\sqsupseteq^{\mathsf{m}}_{\mathsf{lex}}$ the *lexicographic meta-ranking*. Its strict counterpart is denoted by $\sqsupset^{\mathsf{m}}_{\mathsf{lex}}$.

Let us now apply this to our social choice scenario $\mathfrak{C}_2$. On the ex ante reading, when we ask how alternatives a and b compare, we are comparing the ordered tuples $\overrightarrow{D_{\mathsf{e}}(a)}$ and $\overrightarrow{D_{\mathsf{e}}(b)}$, using the lexicographic meta-ranking. In other words, we first look at an individual that is ex ante worst-off given a and an individual that is ex ante worst-off given b. As explained above, these individuals are j, respectively i. Next, we ask whether $\overrightarrow{d_j(a)} \sqsupset_{\mathsf{lex}} \overrightarrow{d_i(b)}$ or whether the converse holds. In our example, we see that in fact, $\overrightarrow{d_j(a)} \sqsupset_{\mathsf{lex}} \overrightarrow{d_i(b)}$ and hence a ranks higher than b.

In case there is a tie in terms of the worst-off individuals, we need to compare the second worst-off. For instance, note that $\overrightarrow{d_j(a)} = \overrightarrow{d_i(d)}$. So to see whether a is better than d or conversely, we need to compare $\overrightarrow{d_i(a)}$ and $\overrightarrow{d_j(d)}$. Since $\langle 2,4 \rangle \sqsupset_{\mathsf{lex}} \langle 2,3 \rangle$, we find that d is better than a.

The following definition covers the procedure in general:

[20] As with tuples of real numbers, we preserve repetitions in tuples of tuples, so that e.g. $\overrightarrow{\langle \langle 1,2 \rangle, \langle 2,4 \rangle, \langle 1,2 \rangle \rangle} = \langle \langle 1,2 \rangle, \langle 1,2 \rangle, \langle 2,4 \rangle \rangle$.

[21] This is essentially just the standard definition of a lexicographic ranking to tuples of tuples, but with $\sqsupseteq_{\mathsf{lex}}$ as the underlying ranking of tuples.

Table 1 Calculations for the four lexical social choice rules, applied to our running example $\mathfrak{C}_2$ in Fig. 6

	Basic	Ex ante	Atomistic	Holistic
a	1, 2, 3, 4	$\langle 1,4\rangle, \langle 2,3\rangle$	$\langle 1,3\rangle, \langle 2,4\rangle$	$\langle 1,2\rangle, \langle 3,4\rangle$
b	1, 3, 3, 3	$\langle 1,3\rangle, \langle 3,3\rangle$	$\langle 1,3\rangle, \langle 3,3\rangle$	$\langle 1,3\rangle, \langle 3,3\rangle$
c	1, 2, 3, 5	$\langle 1,2\rangle, \langle 3,5\rangle$	$\langle 1,2\rangle, \langle 3,5\rangle$	$\langle 1,5\rangle, \langle 2,3\rangle$
d	1, 2, 4, 4	$\langle 1,4\rangle, \langle 2,4\rangle$	$\langle 1,4\rangle, \langle 2,4\rangle$	$\langle 1,2\rangle, \langle 4,4\rangle$
	$b \succ d \succ c \succ a$	$d \succ a \succ b \succ c$	$d \succ b \succ a \succ c$	$c \succ b \succ d \succ a$

The tuples in the second to fifth column represent the way the value of each alternative is represented for these rules; for reasons of space, outer tuple brackets are omitted. The bottom row represents the overall welfare ranking obtained by the respective rules.

Definition 9 (*Ex ante lexical difference rule*) Where $\mathfrak{C} = \langle N, A, S, d\rangle$ is a choice scenario and $a \in A$, a is admissible on the Ex Ante Lexical Difference Rule if and only if there is no $b \in A$ such that $\overrightarrow{D_{\mathsf{e}}(b)} \sqsupset_{\mathsf{lex}}^{\mathsf{m}} \overrightarrow{D_{\mathsf{e}}(a)}$.

Using the running example, we see that only d is admissible on the Ex Ante Lexical Difference Rule. The reasoning leading up to these conclusions is summarized in the second column of Table 1 on page 20.

The atomistic ex post approach When we take an ex post perspective on the distribution of welfare, and if we no longer restrict the focus to the worst-off, then two distinct approaches are possible. On the one hand, one may still focus on the "worst-off, second worst-off, etc." and explicate each of these notions in the ex post sense; after which one then chooses alternatives in view of how they affect those (abstract) persons. We call this the *atomistic (ex post) approach.* On the other hand, one may rather focus on the overall distributions of welfare at outcomes, and compare alternatives in terms of those distributions they leave open. We call this the *holistic (ex post) approach.* We will now explain these two approaches in turn.

On the atomistic approach, we reason as follows. Given any couple $(a, s) \in A \times S$, we know who is (among the) worst-off, second worst-off, and so on in the ex post sense. For instance, in our scenario $\mathfrak{C}_2$, given alternative a and state s_2, i is worst-off (getting only a welfare level of 3), whereas given a and s_1, j is worst-off (getting only a welfare level of 1).

Let us refer to the corresponding welfare levels by $d_{\mathsf{w}:k}(a, s)$ for any $k \leq |N|$. Formally, $d_{\mathsf{w}:k}(a, s)$ is the k^{th} entry in the ordered tuple $\overrightarrow{d(a, s)}$. Similarly, let $\overline{d_{\mathsf{w}:k}(a)} = \langle d_{\mathsf{w}:k}(a, s)\rangle_{s \in S}$. The tuple $\overline{d_{\mathsf{w}:k}(a)}$ represents the welfare of the (ex post) k-worst-off individual given a. In our running example, $\overline{d_{\mathsf{w}:1}(a)} = \langle 1, 3\rangle$ and $\overline{d_{\mathsf{w}:2}(a)} = \langle 2, 4\rangle$. These tuples represent respectively the welfare of the ex post worst-off individual given a, and the welfare of the ex post second worst-off individual given a.

We can now represent the overall value of a given alternative a as the tuple $\overline{D_{\mathsf{w}}(a)} = \langle \overrightarrow{d_{\mathsf{w}:k}(a)}\rangle_{1 \leq k \leq |N|}$. Note that this is, again, a tuple of tuples. Once there, we can again compare alternatives in terms of those tuples, using the lexicographic meta-ranking $\sqsupseteq^{\mathsf{m}}_{\mathsf{lex}}$:

Definition 10 (*Atomistic lexical difference rule*) Where $\mathfrak{C} = \langle N, A, S, d \rangle$ is a choice scenario and $a \in A$, a is admissible on the Atomistic Ex Post Lexical Difference Rule if and only if there is no $b \in A$ such that $\overrightarrow{D_{\mathsf{w}}(b)} \sqsupset^{\mathsf{m}}_{\mathsf{lex}} \overrightarrow{D_{\mathsf{w}}(a)}$.

Applying the running example once more, we obtain the representations of the alternatives depicted in the third column of Table 1 on page 20. Applying the lexicographic meta-ranking to these tuples, we find that d ranks highest, above b, a, and c consecutively. Consequently, d is the only admissible option on the rule just defined. Note that in comparing a and b, we need to consider the second worst-off individuals to break the tie.

The holistic ex post approach While the holistic approach can be instantiated in many different ways—depending on how one evaluates the fairness of single distributions—we here spell it out in the spirit of the leximin rule.

First, we identify the fairness of the welfare distribution at a given outcome (a, s) with the ordered tuple $\overrightarrow{d(a,s)}$. Moving from single outcomes to entire alternatives, this gives us a tuple of such tuples, viz. $\overline{D_{\mathsf{s}}(a)} = \langle \overrightarrow{d(a,s)} \rangle_{s \in S}$. Intuitively, if we order the latter tuples lexicographically, then we know what is the (leximin) "worst possible outcome" given a, the (leximin) "second worst possible outcome" given a, and so on. Finally, we compare the alternatives, again using the lexicographic meta-ranking $\sqsupseteq^{\mathsf{m}}_{\mathsf{lex}}$. In sum:

Definition 11 (*Holistic lexical difference rules*) Where $\mathfrak{C} = \langle N, A, S, d \rangle$ is a choice scenario and $a \in A$, a is admissible on the Holistic Lexical Difference Rule if and only if there is no $b \in A$ such that $\overrightarrow{D_{\mathsf{s}}(b)} \sqsupset^{\mathsf{m}}_{\mathsf{lex}} \overrightarrow{D_{\mathsf{s}}(a)}$.

In our running example, we find that c ranks highest and hence is the only admissible alternative on the Holistic Lexical Difference Rule, since its $\sqsupset_{\mathsf{lex}}$ -worst possible outcome is strictly $\sqsupset_{\mathsf{lex}}$-better than all any $\sqsupset_{\mathsf{lex}}$-worst possible outcome of the other alternatives. Here again we find that, to compare some of the alternatives—i.c. d and a—we need to look at second worst possible outcomes and compare them lexicographically (cf. Table 1).

Taking a step back, we can easily observe that each of the lexical social choice rules are logically distinct. In particular, for our running example, each of the approaches yields a distinct total ranking of the four options a–d (cf. the last row of Table 1). From this it follows that the admissibility verdicts of the various rules are also in general incomparable.

4.2 The impossibility of an original position argument

As noted before, the basic social choice rule $\mathbf{S}^{\mathsf{b}}(\textit{leximin})$ can be grounded in an original position argument, using the leximin rule for individual choice. We now argue that there is no similar original position argument to ground the Ex Ante

$\mathfrak{C}_{\mathrm{d}}$	s_1	s_2
a	$(0,1)$	$(0,1)$
b	$(0,1)$	$(1,0)$
c	$(0,0)$	$(1,1)$

$\mathfrak{C}^*_{\mathrm{d}}$	$(s_1, \pi_=)$	$(s_2, \pi_=)$	$(s_1, \pi_{\neq})$	$(s_2, \pi_{\neq})$
a	$(0,1)$	$(0,1)$	$(1,0)$	$(1,0)$
b	$(0,1)$	$(1,0)$	$(1,0)$	$(0,1)$
c	$(0,0)$	$(1,1)$	$(0,0)$	$(1,1)$

Fig. 7 The choice scenario $\mathfrak{C}_{\mathrm{d}}$ and its OP-transformation $\mathfrak{C}^*_{\mathrm{d}}$

Table 2 The different social choice rules applied to $\mathfrak{C}_{\mathrm{d}}$

	Ex Ante	Atomistic	Holistic
a	−	−	+
b	+	−	+
c	+	+	−

Lexical Rule, the Atomistic Lexical Rule, or the Holistic Lexical Rule. Our argument proceeds in terms of a single choice scenario that is presented in Mongin and Pivato (2021), and goes back to earlier work in the context of social choice under risk (Ben-Porath et al. 1997; Gajdos and Maurin 2004; Diamond 1967).[22] It involves two individuals *i* and *j*, two states s_1 and s_2, and three alternatives *a*, *b*, and *c*, as depicted in Fig. 7 (scenario $\mathfrak{C}_{\mathrm{d}}$). Note that we only need two welfare levels, viz. 0 and 1 in this example.

We first check the recommendations for each of the three mentioned lexical rules in $\mathfrak{C}_{\mathrm{d}}$. On the ex ante reading, the worst-off individual for *a* is individual *i*; this individual gets welfare level 0 twice. In contrast, in alternatives *b* and *c*, the individuals are equally well-off ex ante, getting either 0 or 1. The upshot is that *a* is not admissible on the ex ante reductive account, while both *b* and *c* are.

On the atomistic reading, the worst-off individual given either *a* or *b* will have welfare level 0 at both states. Only with alternative *c* does the worst-off individual have at least some state at which it gets welfare level 1. Consequently, *c* is preferred on the ex post reductive account, while both *a* and *b* are inadmissible.

Finally, on the holistic account, both *a* and *b* are better than *c*. That is, *c* allows for an outcome with distribution (0, 0), whereas *a* and *b* will guarantee that at least one of both individuals get welfare level 1.

Table 2 summarizes our findings up to this point. Importantly, the three social choice rules give distinct recommendations and all three of them rule out at least one of the alternatives. Hence, no individual choice rule that considers all alternatives in $\mathfrak{C}^*_{\mathrm{d}}$ admissible can ground these social choice rules.

[22] These authors use (variants of) this example to argue that simple ex ante or ex post rules give distinct results, and may fail to single out the most "fair" alternative in cases of choice under risk. Here, we use the example for a somewhat different purpose, viz. to show that our three lexical choice rules cannot be reduced to individual choice rules in combination with the OP transformation. Diamond's (1967) original example was targeted against the sure-thing principle as used in Harsanyi's argument for utilitarianism. In turn, it spawned a whole literature on the concept of fairness, cf. Broome (1984, 1990).

$\mathfrak{C}_i$	s_1	s_2	s_3	s_4
a	0	0	1	1
b	0	1	1	0
c	0	1	0	1

$\mathfrak{C}_i^*$	$(s_1, \pi_=)$	$(s_2, \pi_=)$	$(s_3, \pi_=)$	$(s_4, \pi_=)$
a	0	0	1	1
b	0	1	1	0
c	0	1	0	1

Fig. 8 The individualistic choice scenario $\mathfrak{C}_i$ and its OP-transformation $\mathfrak{C}_i^*$

Consider now the choice scenario $\mathfrak{C}_i$ in Fig. 8 with one individual $i \in N$ and its OP-transformation $\mathfrak{C}_i^*$.

Since $\mathfrak{C}_i$ is an individual choice scenario, each of the three lexical social choice rules gives the same recommendation as leximin. Hence, it is plain to see that each of the lexical social choice rules considers every alternative admissible in $\mathfrak{C}_i$.

We now argue by *reductio* that no individual choice rule can exist that grounds one of the three social choice rules in question. Let **S** be the ex ante, atomistic, or holistic lexical rule. Suppose that **R** is an individual choice rule that grounds **S**. Hence, an alternative $x \in \{a, b, c\}$ is **S**-admissible in $\mathfrak{C}_{\mathsf{d}}$ if and only if x is **R**-admissible for i in $\mathfrak{C}_{\mathsf{d}}^*$. Similarly, x is **S**-admissible in $\mathfrak{C}_i$ if and only if x is **R**-admissible for i in $\mathfrak{C}_i^*$. By column symmetry and individualism, x is **R**-admissible for i in $\mathfrak{C}_{\mathsf{d}}^*$ if and only if x is **R**-admissible for i in $\mathfrak{C}_i^*$.[23] But since **R** grounds **S** this must mean that the set of admissible alternatives according to **S** is the same in $\mathfrak{C}_{\mathsf{d}}$ and $\mathfrak{C}_i$. This contradicts our earlier observation that **S** has distinct recommendations for $\mathfrak{C}_{\mathsf{d}}$ and $\mathfrak{C}_i$.

5 Concluding remarks

Let us take a step back and reconsider the original motivation for social choice rules, i.e., to specify when a certain choice is *fair* or *just*. Which of the discussed social choice rules are fair, and which are not? And what does this tell us about original position arguments, as we formalized them?

This is not the place to give ultimate answers to the first question—if such answers exist at all. Focusing on the second, we can however make a number of significant observations. First, if one sticks to the Rawlsian version of the Difference Principle and its emphasis on the least advantaged, then our observations in Sect. 3.2 show that one can easily have a social choice rule that is grounded in an original position argument. In particular, this holds if we favor the basic approach to social choice under ignorance in combination with maximin, if we combine the ex ante approach with a pessimistic notion of "ex ante welfare", or if we combine the ex post approach with maximin. If, in contrast, we deviate from some of these choices—e.g. incorporating a lexical notion of ex ante welfare, or using the leximin rule in the ex post approach, no original position argument can be given. In sum,

[23] We can simply relabel the state labels in $\mathfrak{C}_{\mathsf{d}}^*$ to be in accordance with scenario $\mathfrak{C}_i^*$ and the payoffs of individual i are identical in both scenarios.

only a very narrow conception of the Difference Principle can be grounded in an original position argument.

Second, if one follows Parfit and Van Parijs (and many others) in advocating a Lexical Difference Principle, then our observations in Sect. 4.2 show that the only social choice rule that can be grounded in an original position argument is the basic approach in combination with leximin. However, by its very definition, that rule does not take into account (i) the way distinct individuals are affected by a given choice, and (ii) the distribution of welfare across particular outcomes. This means that, in particular, it cannot distinguish between the three alternatives a, b, c in the example we used in Sect. 4.2.

It can be argued that this clashes with basic egalitarian intuitions that many would associate with fairness. First, a could be deemed inferior to b and c in terms of *ex ante equality*, in that a makes individual i definitely worse off than individual j, no matter what happens. Also, b could be considered inferior to c in terms of *ex post equality*, since c guarantees equality of welfare in every possible state. This type of argument has been used in the context of choice under risk (Mongin and Pivato 2021), but it applies just as well to choices under ignorance. If one agrees with these claims, then this casts serious doubts on the viability of the original position as a way to ground a fair social choice rule.

Does that mean that the three lexical rules we defined in Sect. 4.1 are flawless as standards of fairness? We do not think so. Note first that none of these three rules will give the exact preference relation $c \succ b \succ a$ that is required by the above argument. At best, a combination of these approaches—e.g. first selecting in terms of ex ante welfare, next in terms of ex post equality—could do the job. Moreover, there seems to be a trade-off between the intuition that favours ex post equality and another basic intuition in this type of cases, which we would dub *ex post lexical optimality*. That is, one may argue that in some sense c is worse than both a and b. If the true state of affairs turns out to be s_1, then both individuals may complain: if you had taken alternative a or b, then at least one of them would have been better off and no one would have been worse off. A pessimistic social planner may well try to avoid such a situation, and go for a or b instead.

We do not take a stance for or against these intuitions here. We just note that none of them can be accommodated by any social choice rule that is derived from the original position in combination with an individual choice rule. In this sense, we think that the original position, in the way we have modeled it—following the standard approach to choice under ignorance—cannot account for these basic principles. Either we should allow for richer models of choice under ignorance, or we should give up all three principles just mentioned, or we should give up the idea that social choice rules ought to be grounded in an original position argument.

Acknowledgements Thijs De Coninck is a PhD fellow of the Research Foundation–Flanders supported by a fundamental research Grant (1167619N). Frederik Van De Putte's research was funded by a grant from the Research Foundation-Flanders (FWO-Vlaanderen), no. 12Q1918N, and by a grant from the Dutch Research Council (NWO), No. VI.Vidi.191.105. We are indebted to Allard Tamminga, Stefan Wintein, Diderik Batens, Marcus Pivato, Nathan Wood, Pawel Pawlowski, Marina Uzunova, Federico Faroldi, and three anonymous referees for helpful comments on previous versions of this paper.

References

Ben-Porath, E., Gilboa, I., & Schmeidler, D. (1997). On the measurement of inequality under uncertainty. *Journal of Economic Theory, 75*(1), 194–204.

Bovens, L. (2015). Concerns for the poorly off in ordering risky prospects. *Economics and Philosophy, 31*(3), 397–429.

Broome, J. (1990). Fairness. In *Proceedings of the Aristotelian society* (Vol. 91, pp. 87–101).

Broome, J. (1984). Uncertainty and fairness. *The Economic Journal, 94*(375), 624–632.

Buchak, L. (2017). Taking risks behind the veil of ignorance. *Ethics, 127*(3), 610–644.

Chung, H. (2020). Rawls's self-defeat: A formal analysis. *Erkenntnis, 85,* 1169–1197.

De Coninck, T., & Van De Putte, F. (2021). The original position: A logical analysis. In L. Fenrong, A. Marra, P. Portner, & F. Van De Putte (Eds.), *Deontic logic and normative systems: Proceedings of DEON2020/2021*. College Publications.

Diamond, P. A. (1967). Cardinal welfare, individualistic ethics, and interpersonal comparison of utility: Comment. *The Journal of Political Economy, 75*(5), 765.

Fleurbaey, M. (2018). Welfare economics, risk and uncertainty. *Canadian Journal of Economics, 51*(1), 5–40.

Gajdos, T., & Kandil, F. (2008). The ignorant observer. *Social Choice and Welfare, 31*(2), 193–232.

Gajdos, T., & Maurin, E. (2004). Unequal uncertainties and uncertain inequalities: An axiomatic approach. *Journal of Economic Theory, 116*(1), 93–118.

Gaus, G., & Thrasher, J. (2015). Rational choice and the original position: The (many) models of Rawls and Harsanyi. In T. Hinton (Ed.), *The original position*. Cambridge: Cambridge University Press.

Gustafsson, C. J. E. (2018). The difference principle would not be chosen behind the veil of ignorance. *The Journal of Philosophy, 115*(11), 588–604.

Harsanyi, J. C. (1953). Cardinal utility in welfare economics and in the theory of risk-taking. *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. *American Political Science Review, 69*(2), 594–606.

Harsanyi, J. C. (1977). *Rational behaviour and bargaining equilibrium in games and social situations*. Cambridge: Cambridge University Press.

Hayashi, T., & Lombardi, M. (2019). Fair social decision under uncertainty and belief disagreements. *Economic Theory, 67*(4), 775–816.

Arrow, K. and Hurwicz, L. (1972). An optimality criterion for decision-making under ignorance. In D. F. Carter & F. Ford (Eds.), *Uncertainty and expectations in economics: Essays in Honour of G. L. S. Shackle*. Oxford: B. Blackwell.

Maskin, E. (1979). Decision-making under ignorance with implications for social choice. In *Game theory, social choice and ethics* (pp. 319–337). Springer.

Milnor, J. W. (1954). Games against nature. In R. M. Thrall, C. H. Coombs, & R. L. Davis (Eds.), *Decision processes*. New York: Wiley.

Moehler, M. (2018). The Rawls–Harsanyi dispute: A moral point of view. *Pacific Philosophical Quarterly, 99*(1), 82–99.

Mongin, P., & Pivato, M. (2021). Rawls's difference principle and maximin rule of allocation: A new analysis. *Economic Theory, 71*(4), 1499–1525.

Moreno-Ternero, J. D., & Roemer, J. E. (2008). The veil of ignorance violates priority. *Economics and Philosophy, 24*(2), 233.

Parfit, D. (1991). *Equality or priority*. University of Kansas, Department of Philosophy

Peterson, M. (2017). *An introduction to decision theory*. Cambridge: Cambridge University Press.

Rawls, J. (1971). *A theory of justice*. Harvard: Harvard University Press.

Rawls, J. (1974). Reply to Alexander and Musgrave. *The Quarterly Journal of Economics, 88*(4), 633–655.

Resnik, M.D. (1987). *Choices: An introduction to decision theory*. University of Minnesota Press, Minnesota.

Roemer, J. E. (2002). Egalitarianism against the veil of ignorance. *The Journal of Philosophy, 99*(4), 167–184.

Sen, A. (1970). *Collective choice and social welfare*. San Francisco: Holden Day.

Stefánsson, H.O.(2019). Ambiguity aversion behind the veil of ignorance. *Synthese*, 1–24.

Strasnick, S. (1976). Social choice and the derivation of Rawls's difference principle. *The Journal of Philosophy, 73*(4), 85–99.

Van Parijs, P. (2001). Difference principles. In S. Freeman (Ed.), *The Cambridge companion to John Rawls*. Cambridge: Cambridge University Press.