Presumption of Equality as a Requirement of Fairness

Rabinowicz, Wlodek

Published in:
Descriptive and Normative Approaches to Hyman Behavior

2011

Presumption of Equality as a Requirement of Fairness

Wlodek Rabinowicz
Department of Philosophy, Lund University
wlodek.rabinowicz@fil.lu.se

ABSTRACT: Presumption of Equality enjoins that individuals be treated equally in the absence of discriminating information. My objective in this paper is to make this principle more precise, viewing it as a norm of fairness, in order to determine why and under what conditions it should be obeyed.

Presumption norms are procedural constraints, but their justification might come from the expected outcomes of the procedures they regulate. This outcome-oriented approach to fairness is pursued in the paper. The suggestion is that in the absence of information that would discriminate between the individuals, equal treatment minimizes the expected unfairness in the outcome. Another suggestion is that, under these circumstances, equal treatment also minimizes maximal possible unfairness, i.e., it is minimally unfair if 'worst comes to worst'. Whether these suggestions are correct depends on the properties of the underlying unfairness measure.

1. Introduction

This paper examines Presumption of Equality (PE), which enjoins us to treat different individuals equally if we can't discriminate between them on the basis of the available information. I will view this principle as a requirement of fairness – more specifically, as a procedural principle whose goal is to promote *fairness in outcome*. The objective is to make PE so understood more precise and to determine why and under what conditions it should be obeyed.

Why, then, should PE be obeyed? A natural answer is that, in the absence of relevant discriminating information, treating some individuals better than others is *arbitrary*, which is a bad thing. There's certainly some truth in this explanation. But while arbitrariness considerations are important, they are not always decisive. In some cases in which the discriminating information is absent, unequal treatment might still be right, despite its arbitrariness. To illustrate, suppose two individuals compete for two scholarships, one of which is more attractive than the other. Your task is to make the decision, but the information you have is limited. While you know that both candidates are deserving, you have no clue which of them, if any, has stronger merits. Suppose you have no opportunity to gather further information. Since the scholarships aren't equally attractive, to give one to one individual and the other to the other is to treat them unequally, which is arbitrary and to that extent unsatisfactory. At the same time, one *could* treat them equally by withholding scholarships from both. Such 'levelling down', however, would be grossly unfair given that both of them deserve a scholarship. Avoidance of arbitrariness is thus not all that matters. To justify equal treatment, in cases in which such treatment *can* be justified, we need to rely on other considerations.

At this point, I expect an objection: Why not decide who gets what scholarship by a toss of a coin? This would give each individual a fifty-fifty chance of getting the more attractive scholarship, but even the loser would not come away empty-handed: She would receive the other scholarship instead. Arguably, such a lottery is itself a form of equal treatment, since it gives each person equal chances. By tossing a coin, we avoid arbitrariness but at the same time see to it that both persons are awarded, which they deserve.

It is true that, in the case at hand, drawing lots or tossing a coin is the obvious thing to do. Avoidance of arbitrariness *is* important. I would deny, however, that equal lottery on unequal treatments is on a par with equal treatment. On an outcome-oriented approach to fairness, which is adopted in this paper, this is not so. The outcome of the lottery will still be unequal. And inequality in outcome may well matter from the point of view of fairness.

Here is the suggestion I instead want to examine: Principles of fairness can be constraints on procedures or constraints on outcomes. Presumption norms such as PE constrain procedures, but procedural constraints can often be justified in terms of the expected outcomes of the procedures that obey these constraints. This is the path that will be pursued in my paper: The suggestion is that equal treatment should be chosen because, and to the extent that, it *minimizes the expected unfairness* in outcome. When the available information does not discriminate between the individuals concerned, expected unfairness will normally be at its lowest if individuals are treated equally. Normally, but not always. As will be seen, the scholarship example provides an exception to this rule.

It has been suggested that the core reason behind presumption norms is to be found in the differential costs of potential errors.[1] Thus, for example, presumption of innocence in criminal law is justified by the greater moral cost of punishing an innocent person as compared with that of letting a guilty person go free. Louis Katzner applied this idea to the choice between presumption of equality and its opposite, presumption of inequality:

> The only possible basis for opting for one of them rather than the other is which state of affairs one would rather see – that in which some of those who are similar are treated differently or that in which some of those who are different are treated similarly. (Katzner 1973, p. 92)[2]

My approach to PE is different. As will be seen, for this principle to hold, the moral cost of treating equals unequally need not be greater than that of the equal treatment of unequals.

Some presumptions might have more to do with the differential probability of errors than with their differential costs: What's being presumed is deemed to be sufficiently probable to function as a default assumption.[3] But my argument

---

[1]See Ullmann-Margalit (1983), p. 159: "It is the justification of presumptions in normative terms which touches what I take to be the core of the concept of presumption. […] this normative type of consideration has to do with the acceptabil- ity of error." (p. 159) Making a presumption is grounded in "certain evaluative considerations which are primarily concerned with the differential acceptability of the relevant sorts of expected errors: the fact that one sort of error is judged to be, in the long run and all things considered, preferred on grounds of moral values or social goals to the alternative sort(s) constitutes an overriding reason for the decision underlying the presumption rule." (p. 162)

[2] Quoted in Ullmann-Margalit (1983).

[3] Cf. Ullmann-Margalit (1983), p. 157: "with presumption rules relating to presumptions that accord with the normal balance of probability the chance of an error […] is reduced."

does not assume that equality in deserts is probabilistically privileged in this way. Indeed, PE might well be justified even in those cases in which it is very *im*probable, or perhaps even excluded, that the individuals are equally deserving.

While the justification I offer does not appeal to the differences in the costs of errors or in the probabilities of errors, it does appeal to the differences in the *expected* costs of errors. The suggestion is that in the absence of discriminating information we should treat individuals as if they were equal because the expected moral cost of error is minimized in this way: Equal treatment minimizes expected unfairness. To get an idea why it is so, note that unfairness may be seen as a kind of distance between the way individuals are treated and the way they deserve to be treated. A treatment's expected unfairness is on this view its expected distance from the fair treatment. We can thus think of the set of possible treatments as a set of points forming a spatial area. One of these points is the fair treatment but, in the absence of information, we don't know which point it is. Consequently, we don't know how large is the distance between the treatment we choose and the fair treatment. Now, the conjecture is that equal treatment lies in the center of this spatial area. At the same time, in the absence of information that discriminates between individuals, no dírection in the area is more privileged than other directions. Therefore, by positioning ourselves in the center we should minimize our *expected* distance from the fair treatment, i.e. keep the expected unfairness in outcome at a minimum.

I want to examine under what conditions on the unfairness measure equal treatment will in fact have this feature of centrality. I will also inquire what happens if we instead choose a 'minimax' approach, i.e. opt for a treatment that is least unfair if 'worst comes to worst'. In other words, I will also examine under the conditions on the unfairness measure under which equal treatment minimizes maximal possible unfairness. Intuitively, centrality should guarantee this result as well.

2. Individuals and Treatments

The model to be used is highly abstract and allows of different interpretations. Its main components are a non-empty finite set **I** = $\{i_1, ..., i_n\}$ of *individuals* and a non-empty set **T** = $\{a, b, c, ...\}$ of possible *treatments* of individuals in **I**.

Every treatment a in **T** is assumed to be a vector $(a_1, \ldots, a_n)$, where, for every k $(1 \leq k \leq n)$, a_k is the way in which individual i_k is treated in a. We shall sometimes use the notation $a(i_k)$ for a_k. Treatment a is *equal* iff $a_1 = \ldots = a_n$. Other components of the model will be introduced later.

The three interpretations of the model that follow are themselves relatively abstract. Each can in turn be instantiated in many different ways.

Interpretation 1: Cake-Divisions

A 'cake' is a homogeneous object or resource that is to be divided, without remainder, among the individuals in **I**. A treatment a is a vector of real numbers, $(a_1, ..., a_n)$, with each a_k being the share of the cake assigned by a to individual i_k. The shares are all non-negative and together they sum up to one. **T** is the set of all possible vectors of this kind. The equal treatment, $(1/n, ..., 1/n)$, divides the cake equally among the members of **I**.

Representing cake-divisions in this way means that we view them as *types* rather than tokens. Thus, to illustrate, if a cake is divided in pieces of equal size, it doesn't matter who gets which piece, since the cake is homogeneous. There is therefore no reason to make this distinction in the model. This is a general feature of our approach. Treatments are interpreted as types that specify the relevant characteristics of their tokens. As a result, any two treatments in the model are supposed to be relevantly different from each other.

Interpretation 2: Rankings

On this interpretation, **T** is the set of all possible rankings of the individuals in **I**. That a treatment a ranks i above j means that i is treated better than j in a. A tie in a between i and j means that they are treated equally well. The ranking interpretation of treatments is appropriate when *ordinal* differences between the individuals are all that matters from the point of view of fairness, i.e., when fairness only requires that the more deserving individuals should be better treated and that the equally deserving individuals should be treated equally well.

A ranking may be represented as an assignment of ordinal numbers to individuals, with 1 being the highest level in the ranking, 2 being the second highest level, etc. The assignment of levels starts from the highest one and continues downwards. Thus, equal treatment is the ranking in which every individual is assigned the highest level: (1, …, 1). (For another logically equivalent representation of rankings see below, section 3.)

Interpretation 3: Indivisible Goods
Suppose that G is a set of indivisible objects that are to be distributed, with or without remainder, among the individuals in **I**. $a(i)$ is the subset of G that treatment a assigns to an individual i. For some i, $a(i)$ may be empty, and for distinct i and j, $a(i)$ and $a(j)$ are disjoint, i.e. have no elements in common. Some objects in G may be withheld from the distribution. The scholarship case provides an example. There, G consists of two scholarships, the more attractive one, A, and the less attractive B, and possible treatments amount to different partial or total distributions of G among the two individuals involved. Thus, for example, $a = (\{A\}, \{B\})$ is the assignment in which A goes to i and B goes to j, while $b = (\{B\}, \{A\})$ assigns B to i and A to j. The equal treatment is the distribution $(\varnothing, \varnothing)$ in which scholarships are withheld from both idividuals, i.e., each of them is assigned the empty set.

For simplicity, I exclude decision problems in which there is no equal treatment or in which several treatments are equal. (The latter restriction would be violated, for example, in Cake-Division, if we allowed divisions in which part of the cake remains undistributed. The number of equal treatments would then increase from one to infinity.) I will also assume that **T** is closed under permutations on individuals. Thus, we impose two conditions on the set of treatments:

A1. For every permutation f on **I** and every a in **T**, **T** contains some b such that for every i in **I**, $f(i)$ is treated in b as i is treated in a. I.e., for every i, $b(f(i)) = a(i)$.

A2. There is a unique element of **T**, call it **e**, such that **e** is an equal treatment.

A2 is a substantial restriction. So is A1, which is a kind of completeness requirement on **T**. If **T** is the set of actually *available* treatments, this set sometimes might be too small for A1 to be satisfied. Here, however, I will ignore this difficulty and assume that **T** is ‘roomy’ enough.

There are two ways of looking at **T**. If **T** is seen as the set of *conceivable* treatments, it is plausible to suppose that **T** is large enough to satisfy such conditions as A1. This way of looking at **T** is appropriate if we think of the elements of **T** as the possible ways in which the individuals might *deserve* to be treated. But if **T** instead is interpreted as the set of ways in which we can treat the individuals, i.e., as the set of *available* treatments, A1 is not that plausible. Still, in the context of our discussion, the simplifying assumption that the set of available treatments is as large as the set of conceivable treatments is innocuous: Remember that we want to know whether equal treatment minimizes expected unfairness (and/or maximal possible unfairness), as compared with available alternatives, in the absence of information that discriminates between the individuals. If this conjecture turns out to hold when the set of available treatments is large, it will obviously still hold if that set is diminished.

Given A1, every permutation *f* on **I** induces the corresponding permutation on **T** that for to every *a* in **T** assigns some *b* in **T** in which for every individual *i*, *f*(*i*) is treated in the same way as *i* is treated in *a*. I will refer to the union of *f* and the permutation that *f* induces on **T** as an *automorphism* and use symbols *p*, *p'*, etc, to stand for different automorphisms. Intuitively, then, an automorphism is a simultaneous permutation of individuals into individuals and of treatments into treatments, in which the former permutation induces the latter.

> *Definition:* An *automorphism*, *p*, is a simultaneous permutation of **I** and of **T** such that for all *i* in **I** and all *a* in **T**, $p(a)(p(i)) = a(i)$.

> *Corollary of A1:* Every permutation on **I** is included in exactly one automorphism.

This notion of an automorphism will come in handy below.

Here follow some examples of authomorphisms. Suppose that **I** consists of three individuals, i_1, i_2, i_3, and let **T** be the set of cake-divisions among the members of **I**. One automorphism would then permute i_1 into i_2, i_2 into i_3 and i_3 into i_1. This would effect the corresponding permutation on cake-divisions. For example, (0, 2/3, 1/3) would be permuted into (1/3, 0, 2/3). Analogously, if **T** is the set of rankings of i_1, i_2, and i_3, the automorphism that permutes i_1 into i_2,

i_2 into i_3 and i_3 into i_1 involves the corresponding permutation on rankings. For example, it permutes the ranking with i_1 on top, followed by i_2 and i_3, in that order, into the ranking with i_2 on top, followed by i_3 and i_1.

It is easy to see that only equal treatment, **e**, stays invariant under all automorphisms: For all p, $p(\mathbf{e}) = \mathbf{e}$, and for all $a \in \mathbf{T}$, if $a \neq \mathbf{e}$, then for p, $p(a) \neq a$.

We now define a relation between treatments that's going to be important in what follows:

> *Structural Identity:* A treatment a is *structurally identical* to a treatment b iff there exists some automorphism p such that $p(a) = b$.

Intuitively, this relation obtains between two treatments if we can get one from the other just by reshuffling individuals, while otherwise keeping the treatment unchanged. Structural identity is an equivalence relation: it is reflexive, symmetric, and transitive.[4] We can therefore partition **T** into *structures*, S, S', etc., which are equivalence classes of treatments with respect to the relation of structural identity. As an example, suppose that **T** is the set of cake-divisions among three individuals, i_1, i_2, and i_3. Consider a cake-division $a = (1, 0, 0)$. Its structure consists of three treatments:

$$(1, 0, 0), (0, 1, 0) \text{ and } (0, 0, 1).$$

On the other hand, the structure of $b = (1/2, 1/3, 1/6)$ consists of six treatments. In b, each of the three individuals gets a different share and there are six ways in which we can assign these three different shares to three individuals.

For any treatment a in **T**, let S_a be the structure of a. The number of treatments in a structure may vary but is always finite given that **I** is finite: If **I** contains n individuals, the number of treatments in a structure is at most equal to $n!$, which is the number of possible permutations on **I**. Different

[4] This follows because it is true by definition that the set of automorphisms contains the identity automorphism and is closed under inverses and relative products.

automorphisms correspond to different permutations on **I** and the number of treatments in a structure cannot exceed the number of automorphisms. But it can be smaller: In some cases, several permutations on **I** induce automorphisms transforming *a* into the same treatment, which decreases the size of S_a. This will be the case whenever two or more individuals are treated equally in *a*. Thus, for example, there are two permutations on individuals that give rise to automorphisms transforming (1, 0, 0) into (0, 1, 0). Each of them assigns i_2 to i_1, but they differ in their assignments to i_2 and i_3. Since the latter two individuals are treated equally in (1, 0, 0), in both cases (1, 0, 0) is transformed into the same treatment, (0, 1, 0).

At one extreme, all the individuals are treated in the same way in **e**. Therefore, **e**'s structure contains only **e** itself. At the other extreme, if all the individuals are treated differently in *a*, any two distinct automorphisms will transform *a* into different treatments, which means that the number of treatments in *a*'s structure will equal the number of automorphisms.

3. Unfairness Measure

Before I introduce the last component of the model, an unfairness measure, let me first make a further simplifying assumption: In the situations to be considered, the agent knows that *there is exactly one (perfectly) fair treatment in* **T**, i.e. exactly one treatment in which everyone gets what he or she deserves. That there is at least one such treatment in **T** is plausible if we think of **T** as the set of coceivable ways in which the individuals might deserve to be treated. (See the discussion of A1 in the preceding section.). But that there is no more than one such way is a non-trivial constraint on the model. In some cases, it would need to be given up.

The unfairness measure **d** is based on this assumption. **d** is a function from pairs of treatments to real numbers, with the following interpretation: **d**(*a*, *b*) specifies the degree of unfairness of *a* on a hypothetical supposition that it is *b* that is the (perfectly) fair treatment. This degree of unfairness can be seen as the *distance* from *a* to *b*: To the extent *a*'s unfairness is greater, *a* is farther away from *b*. On this interpretation, **d** can be assumed to satisfy the standard conditions on a distance measure:

(D0) $\mathbf{d}(a, b) \geq 0$; *(Non-negativity)*

(D1) $\mathbf{d}(a, b) = 0$ iff $a = b$; *(Minimality)*

(D2) $\mathbf{d}(a, b) = \mathbf{d}(b, a)$ *(Symmetry)*

(D3) $\mathbf{d}(a, b) + \mathbf{d}(b, c) \geq \mathbf{d}(a, c)$ *(Triangle Inequality)*

Interpreting **d** as distance gives the model a geometric flavour. The pair (**T**, **d**) is then a *metric space*: a set of points with a distance measure defined on it. The set **I** of individuals may be seen as the set of *dimensions* of that space: A point $a = (a_1, \ldots, a_n)$ is defined by its coordinates on the different dimensions i_k in **I**.

As will be seen in the next section, interpreting unfairness as distance goes well beyond what's needed for our purposes. In particular, it needn't be assumed that the unfairness measure is symmetric or that it satisfies triangle inequality. Especially since some of these superfluous assumptions, such as symmetry, have quite notable implications.[5] Nevertheless, this geometric interpretation is not implausible and, in addition, it makes the model more intuitive and easier to grasp.

One further very natural condition on **d** is Impartiality, which requires **d** to be invariant under automorphisms:

Impartiality: For all automorphisms p and all a, b in **T**, $\mathbf{d}(p(a), p(b)) = \mathbf{d}(a, b)$.

According to this condition, if one permutes the individuals in two treatments in the same way, the distance between the treatments doesn't change. This means that the unfairness measure pays no attention to personal identities. Thus, for example, giving all of the cake to the individual who only deserves a small share is equally unfair independently of who it is who gets this unfair advantage.

---

[5] Thus, symmetry implies that for every a in **T**, $\mathbf{d}(\mathbf{e}, a) = \mathbf{d}(a, \mathbf{e})$, which means that equal treatment of unequals is just as unfair as the correspondingly unequal treatment of equals. This means that, in the presence of symmetry, justification of PE cannot be traced back to the differential moral costs of errors (cf. section 1 above).

How is **d** to be understood on the different interpretations of our model? Consider Cake-Divisions first. It seems plausible that for each individual *i*, the distance between two cake-divisions should be an increasing function of the (absolute) difference between the shares of the cake they give to *i*. I.e., the distance between treatments *a* and *b* should be an increasing function of $|a_1 - b_1|, \ldots , |a_n - b_n|$. The simplest function of this kind is the sum: $|a_1 - b_1| + \ldots + |a_n - b_n|$. This kind of measure is sometimes called *city-block* distance. If we instead go for the sum of the *squared* differences and then take the square root of that sum, we get Euclidean distance. City-block and the Euclidean measure are two instances of the class of *Minkowski distance* functions. **d** belongs to this class iff, for some $k \geq 1$, and for all *a*, *b* in **T**, **d**(*a*, *b*) = the *k*-th root of the sum $|a_1 - b_1|^k + \ldots + |a_n - b_n|^k$. If $k = 1$, the so-defined **d** is the city-block distance; if $k = 2$, it is the Euclidean distance. The higher *k* is, the more disproportionate influence is given, by exponentiation, to the larger differences $|a_i - b_i|$, as compared with the smaller differences. Only if $k = 1$, all the differences are given influence proportionate to size.

How is the distance measure to be understood for Rankings? Again, there are several possibilities, but the proposal due to Kemeny and Snell seems especially plausible. A ranking might be seen as a set of ordered pairs of individuals: A pair (*i*, *j*) belongs to a ranking *a* iff *a* ranks *i* at least as highly as it ranks *j*. We fully specify a given ranking by providing a list of such ordered pairs. Now, the distance between two rankings, *a* and *b*, can be measured by the number of pairs that belong to either *a* or *b* but not to both of these rankings. It is easy to show that this definition satisfies the standard conditions on a distance measure (D0 - D3 above).[6]

For the case of Indivisible Goods we don't have a plausible definition of distance that is *generally* applicable. Different situations that exemplify the

---

[6]Cf. Kemeny & Snell 1962, chapter on *Preference Ranking: An Axiomatic Approach*. The authors show that this measure is the only distance function on rankings that satisfies the following condition:

If a ranking *b* lies 'between' rankings *a* and *c*, in the sense that it is included in their union and includes their intersection, then **d**(*a*, *b*) + **d**(*b*, *c*) = **d**(*a*, *c*).

I.e., if *b* lies between *a* and *c* in the specified sense, then, intuitively, *b* lies on the shortest line connecting *a* and *c*.

general structure of Indivisible Goods require different specifications of the unfairness measure. Let us therefore focus on the scholarship example we have started with. The example involves two individuals, *i* and *j*, and two scholarships, one more attractive, *A*, and the other, *B*, less so. Let us suppose that one of them (though unknown which one) is more deserving than the other.[7]

There is no need to fully specify a suitable unfairness measure **d**. I will only assume that the following holds: the distance between alternative treatments in which each individual gets a scholarship, i.e., the distance between $a = (\{A\}, \{B\})$ and $b = (\{B\}, \{A\})$, is shorter than the distance to each of these treatments from the equal treatment $\mathbf{e} = (\varnothing, \varnothing)$, in which none of the individuals gets anything. To put it formally,

$$\mathbf{d}(\mathbf{e}, a) > \mathbf{d}(b, a) = \mathbf{d}(a, b) < \mathbf{d}(\mathbf{e}, b).$$

Intuitive motivation: If *a* is fair or if *b* is fair, both individuals deserve a scholarship. But then withholding the scholarship from both is even more unfair than giving the somewhat more attractive scholarship to the less deserving individual. Also, since **d** is symmetric, $\mathbf{d}(b, a) = \mathbf{d}(a, b)$.

4. Information Measure and Expected Unfairness

Apart from the fixed components, **I**, **T**, **e**, and **d**, our model contains one variable component: a probability distribution *P* on **T**, which reflects the agent's information about the case at hand. For every *a* in **T**, $P(a)$ stands for the agent's probability for *a* being the (perfectly) fair treatment. [8]

---

[7] This assumption is made in order to guarantee that the fair treatment belongs to **T**. If both candidates were equally deserving, none of the available treatments would be fair.

[8] Fixed components are marked in bold, while the variable component is italicized. We take *P* to be variable, since we want to examine whether equal treatment ought to be chosen for every *P* that does not discriminate between the individuals.

Since, as we have assumed, the agent knows that there is exactly one fair treatment in **T**, we take it that the *P*-values for different treatments sum up to one. There is a difficulty here, though. On some interpretations, such as Cake-Divisions, the number of possible treatments in **T** is infinite. In such cases, the sum of *P*-values for different treatments might be lower than one and might even be zero if the probability is distributed uniformly over **T**. In such a uniform distribution, each treatment gets the probability zero (unless we allow infinitesimals as probability values) and the sum of zeros is zero. This difficulty could be dealt with by replacing summation with integration, but to keep the calculations at the elementary level I will assume that there exists a finite subset of **T** such that the agent is certain that the fair treatment belongs to that subset. Then *P*-values of different treatments will sum up to one, as desired, and, for any subset *Y* of **T**, we can define *P*(*Y*), the probability that *Y* contains the fair treatment, as the sum of *P*-values assigned to the elements of *Y*:

$$P(Y) = \Sigma_{a \in Y} P(a).$$

To say that the available information does not discriminate between the individuals in **I** must mean that structurally identical treatments are assigned the same *P*-values. Thus, we are led to the following definition:

> *P does not discriminate between the individuals* iff for all structurally identical *a* and *b* in **T**, *P*(*a*) = *P*(*b*).[9]

Two treatments are structurally identical iff there is an automorphism that transforms one into the other. Thus, *P* does not discriminate between the individuals iff it is *invariant under automorphisms*: for every *p* and *a*, $P(a) = P(p(a))$. Note also that if *P* does not discriminate between individuals, then for every *a*, *P*(*a*) equals the probability of *a*'s structure S_a, divided by the number of treatments that belong to this structure: $P(a) = P(S_a)/\text{card}(S_a)$.

---

[9] Our finiteness constraint on *P* doesn't hinder *P* from being indiscriminative in this way. The reason is that every structure is finite if **I** is finite.

It is easy to define the *expected unfairness* of a treatment *a* with respect to a given probability function *P* as the *P*-weighted sum of its distances to different possible treatments. For every treatment *b*, the distance from *a* to *b* is weighted with the probability *P*(*b*) of *b* being *the* fair treatment.

Expected unfairness: $\text{ExpUnf}_P(a) = \Sigma_{b \in \mathbf{T}} P(b)\, \mathbf{d}(a, b)$.

For the expected value to be a meaningful notion, it's enough if the underlying value function is unique up to positive affine transformations, i.e., up to the choice of unit and zero. Representing unfairness as *distance* implies that the only thing that's left for an arbitrary decision is the unit of measurement. The zero-point for distance is not arbitrary: That each point's distance to itself, and only to itself, equals zero is a defining feature of a distance measure.

Even apart from the fixity of the zero-point, it should be clear that interpreting the unfairness measure **d** as a distance function is much more than we need in order to give meaning to the notion of expected unfairness. Neither symmetry nor triangle inequality are needed for this purpose. But, as suggested above, treating unfairness as distance is not implausible and it makes the model easier to grasp.

5. Expected Unfairness and Equal Treatment

Consider the following hypothesis:

> *ExpUnf-minimization*: For every *P* that does not discriminate between the individuals and for every treatment *a* in **T**, $\text{ExpUnf}_P(\mathbf{e}) \leq \text{ExpUnf}_P(a)$.

In other words, on this hypothesis, equal treatment minimizes expected unfairness in the absence of discriminating information. This would explain why PE should be accepted.

We want to know under what circumstances the hypothesis is going to hold. More precisely, we want to know what condition on the *unfairness measure* would make ExpUnf-minimization valid.

If Y is a finite set of treatments, let $\bar{\mathbf{d}}(a, Y)$ stand for a's *average distance* to the treatments in Y:

$$\bar{\mathbf{d}}(a, Y) = \Sigma_{b \in Y}\, \mathbf{d}(a, b)/\mathrm{card}(Y).$$

The following condition on **d** can be shown to be both *necessary and sufficient* for ExpUnf-minimization:

Structure Condition: For every structure $S \subseteq \mathbf{T}$ and every $a \in \mathbf{T}$, $\bar{\mathbf{d}}(\mathbf{e}, S) \leq \bar{\mathbf{d}}(a, S)$.

The condition states that, for every structure S, equal treatment has a minimal average distance to S, as compared with other treatments.

Sufficiency: Structure Condition $\Rightarrow$ ExpUnf-minimization

Proof:

Claim: If P does not discriminate between the individuals, then for every $a \in \mathbf{T}$,

$$\mathrm{ExpUnf}_P(a) = \Sigma_{S \subseteq \mathbf{T}}\, P(S)\, \bar{\mathbf{d}}(a, S).$$

I.e., in the absence of discriminating information, a's expected unfairness is a weighted sum of its average distances to different structures, with weights being the probabilities of these structures. Here's the proof of the Claim:

$$\begin{aligned}
\mathrm{ExpUnf}_P(a) &= \Sigma_{b \in \mathbf{T}}\, P(b)\, \mathbf{d}(a, b) && \text{[by the definition of ExpUnf]} \\
&= \Sigma_{S \subseteq \mathbf{T}}\, \Sigma_{b \in S}\, P(b)\, \mathbf{d}(a, b) && \text{[since } \mathbf{T} \text{ can be partitioned into structures]} \\
&= \Sigma_{S \subseteq \mathbf{T}}\, \Sigma_{b \in S}\, (P(S)/\mathrm{card}(S))\, \mathbf{d}(a, b) && \text{[since } P \text{ is indiscriminative]} \\
&= \Sigma_{S \subseteq \mathbf{T}}\, P(S)\, (\Sigma_{b \in S}\, \mathbf{d}(a, b)/\mathrm{card}(S)) && \text{[by algebra]} \\
&= \Sigma_{S \subseteq \mathbf{T}}\, P(S)\, \bar{\mathbf{d}}(a, S) && \text{[by the definition of } \bar{\mathbf{d}}\text{]}
\end{aligned}$$

Given the Structure Condition, the average distance from **e** to a structure *S* never exceeds the corresponding distance from any *a* to *S*. Consequently, Claim implies that $\text{ExpUnf}_P(\mathbf{e}) \leq \text{ExpUnf}_P(a)$. □

We now want to prove that the Structure Condition is *necessary* for ExpUnf-minimization:

Necessity: Structure Condition $\Leftarrow$ ExpUnf-minimization

Proof: We need to show that if the Structure Condition is violated by our model, i.e., if for some structure *S* and treatment *a*, $\bar{\mathbf{d}}(a, S) < \bar{\mathbf{d}}(\mathbf{e}, S)$, then ExpUnf-minimization is violated as well: there exists a probability function *P* that does not discriminate between the individuals and is such that, with respect to *P*, the expected unfairness of *a* is lower than the expected unfairness of **e**. To construct a *P* like this, we simply let it be the uniform probability distribution on *S*. □

To forestall possible misunderstandings, it should be pointed out that for a *particular* *P* that does not discriminate between the individuals, **e** might minimize expected unfairness with respect to that *P* even if the underlying unfairness measure **d** happens to violate Structure Condition. However, Structure Condition is necessary if **e** is to minimize expected unfairness for *all* possible *P* that do not discriminate between the individuals, as required by the hypothesis of ExpUnf-minimization.

Is Structure Condition satisfied by the different interpretations of our model? I think it is fair to say that this condition *usually* holds. It can be shown to hold for all Minkowski-distance measures on cake-divisions.[10] It can also be

---

[10]For the proof, see Rabinowicz (2008), Appendix A. In that appendix, I consider a more general interpretation on which **T** consists of all possible real-number assignments to individuals, i.e., not only those in which the assigned numbers are non-negative and add up to 1 (as in Cake-Divisions). On this interpretation, there are non-denumerably many equal treatments in **T**. The following condition is shown to be satisfied by every Minkowki-distance measure on such a set **T**:

shown to hold for the Kemeny-Snell distance measure on rankings.[11] On the other hand, this condition is violated in the scholarship example. There, as we remember, the distance between treatments $a = (\{A\}, \{B\})$ and $b = (\{B\}, \{A\})$ is shorter than the distance to each of them from the equal treatment $\mathbf{e} = (\emptyset, \emptyset)$. Since the set $\{a, b\}$ is a structure, it immediately follows that the average distance from a to this structure is shorter (in fact, more than twice as short) than the corresponding average distance from **e** to the structure in question:

$$\bar{\mathbf{d}}(a, \{a, b\}) = (\mathbf{d}(a, a) + \mathbf{d}(a, b))/2 = \mathbf{d}(a, b)/2$$
$$< (\mathbf{d}(\mathbf{e}, a) + \mathbf{d}(\mathbf{e}, b))/2 = \bar{\mathbf{d}}(\mathbf{e}, \{a, b\}).$$

Since Structure Condition is violated in this case, it follows that there exists a probability function P that does not discriminate between the individuals and with respect to which **e**'s expected unfairness exceeds the expected unfairness of a: One such P is the uniform probability distribution on $\{a, b\}$. If we are certain that the fair treatment is either a or b, with each of these treatments being an equally likely candidate to the title, treating the individuals equally by withholding the scholarships from both will not minimize expected unfairness.

Generalized Structure Condition: For every $a \in \mathbf{T}$, there exists an equal treatment $\mathbf{e}_a \in \mathbf{T}$ such that for every structure $S \subseteq \mathbf{T}$, $\bar{\mathbf{d}}(\mathbf{e}_a, S) \leq \bar{\mathbf{d}}(a, S)$.

For Minkowski spaces, $\mathbf{e}_a$ is obtained from a by averaging:

For every individual i, $\mathbf{e}_a(i) = (a_1 + \ldots. + a_n)/n$.

When **T** is restricted to the set of Cake-Divisions, in which real values assigned to the different individuals add up to 1, $\mathbf{e}_a$ will coincide with **e**, for every cake-division a. Therefore, for this restricted set of treatments, the simple Structure Condition will hold.

[11] For the proof, see Rabinowicz (2008), Appendix B.

As we have seen in this section, Structure Condition is both sufficient and necessary if equal treatment is to minimize expected unfairness. But this condition is neither especially transparent nor intuitive. It has a feel of a constraint that itself should be derivable from some more basic and simple conditions. What these conditions might be is not clear to me, however. One of them would probably be Impartiality mentioned in section 3, which implies that for each structure, all its elements are equi-distant from the equal treatment. But we obviously need other conditions as well.[12] The conjecture is that they would guarantee, together with Impartiality, that for every structure, the treatments in that structure form vertices of a regular geometric figure that has equal treatment in its center, thereby implying the Structure Condition. Finding these underlying conditions is, as I see it, a major outstanding problem that is left for further inquiry.

6. Minimax

Given Structure Condition, equal treatment minimizes expected unfairness in the absence of discriminating information. However, for some people, this might not be a decisive consideration. They might feel that the proper course of action is not to minimize expected disvalue but rather to minimize the maximal potential disvalue, i.e. to minimize unfairness in *the worst possible case* that has a non-zero probability. What is the position of equal treatment from this 'minimax' perspective?

A treatment *a*'s *maximal possible unfairness* with respect to a probability distribution *P* can be defined as *a*'s maximal distance to a positively *P*-valued treatment:

Maximal possible unfairness:

$$\mathrm{MaxUnf}_P(a) = \max\{\mathbf{d}(a, b): b \in \mathbf{T} \ \&\ P(b) > 0\}.$$ [13]

---

[12] Impartiality is satisfied in the scholarship example despite the fact that this example violates Structure Condition.

[13] If the set $\{\mathbf{d}(a, b): b \in \mathbf{T} \ \&\ P(b) > 0\}$ is infinitely large, it might lack a maximum (or even an upper bound). However, this problem won't arise as long as we hold on

With this notion in hand, we can consider the following hypothesis:

> *MaxUnf-minimization:* For all P that do not discriminate between the individuals, and all a in **T**, $\mathrm{MaxUnf}_P(\mathbf{e}) \leq \mathrm{MaxUnf}_P(a)$.

If this hypothesis holds, it would provide another potential reason for accepting PE.

As is easily seen, the following condition on **d** is both necessary and sufficient for the validity of MaxUnf-minimization:

> *Minimax Condition:* For every a in **T** and every structure $S \subseteq \mathbf{T}$,
> $$\max\{\mathbf{d}(\mathbf{e}, b): b \in S\} \leq \max\{\mathbf{d}(a, b): b \in S\}.$$

According to Minimax Condition, equal treatment minimizes maximal distance to every structure, as compared with other treatments in **T**.

Sufficiency: Minimax Condition $\Rightarrow$ MaxUnf-minimization

Proof: Let c be any treatment such that $P(c) > 0$ and $\mathrm{MaxUnf}_P(\mathbf{e}) = \mathbf{d}(\mathbf{e}, c)$. Let S_c be the structure of c. By Minimax Condition, it holds for every a in **T** that

$$\max\{\mathbf{d}(\mathbf{e}, b): b \in S_c\} \leq \max\{\mathbf{d}(a, b): b \in S_c\}.$$

But then, for some $c' \in S_c$, $\mathbf{d}(a, c') = \max\{\mathbf{d}(a, b): b \in S_c\} \geq \max\{\mathbf{d}(\mathbf{e}, b): b \in S_c\} = \mathbf{d}(\mathbf{e}, c)$. Since $P(c) > 0$ and P does not discriminate between the individuals, $P(c') = P(c) > 0$. Consequently, $\mathrm{MaxUnf}_P(a) \geq \mathbf{d}(a, c')$. Since $\mathbf{d}(a, c') \geq \mathbf{d}(\mathbf{e}, c) = \mathrm{MaxUnf}_P(\mathbf{e})$, it follows that $\mathrm{MaxUnf}_P(a) \geq \mathrm{MaxUnf}_P(\mathbf{e})$. □

Necessity: Minimax Condition $\Leftarrow$ MI-minimization

---

to our simplifying assumption that the number of treatments in **T** with positive P-values is finite.

Proof: Suppose that for some a and S, Minimax Condition is violated:

$$\max\{\mathbf{d}(\mathbf{e}, b): b \in S\} > \max\{\mathbf{d}(a, b): b \in S\}.$$

Let P be the uniform probability distribution on S. Then P does not discriminate between the individuals, but $\text{MaxUnf}_P(\mathbf{e}) = \max\{\mathbf{d}(\mathbf{e}, b): b \in S\} > \max\{\mathbf{d}(a, b): b \in S\} = \text{MaxUnf}_P(a)$. Which means that MaxUnf-minimization is violated as well. □

If the distance measure satisfies the Structure Condition and Impartiality, there is no need to impose Minimax Condition as an independent constraint. It can be shown that

Structure Condition & Impartiality $\Rightarrow$ Minimax Condition.

Proof: Impartiality implies that for every a in $\mathbf{T}$ and every automorphism p, $\mathbf{d}(\mathbf{e}, a) = \mathbf{d}(p(\mathbf{e}), p(a)) = \mathbf{d}(\mathbf{e}, p(a))$. Since a and b are structurally identical iff $b = p(a)$ for some automorphism p, and since $\mathbf{e}$ is invariant under automorphisms, the following must hold given Impartiality:

For all structurally identical treatments a, b in $\mathbf{T}$, $\mathbf{d}(\mathbf{e}, a) = \mathbf{d}(\mathbf{e}, b)$.

Thus, for every structure S, $\mathbf{e}$ is equi-distant from every treatment in S. Now, consider any treatment a in $\mathbf{T}$. By the Structure Condition, $\bar{\mathbf{d}}(\mathbf{e}, S) \leq \bar{\mathbf{d}}(a, S)$. Therefore, since $\mathbf{e}$'s distance to different treatments in S is constant, the maximal distance from $\mathbf{e}$ to the elements of S cannot exceed the maximal distance from a to the elements of S. Minimax Condition follows. □

Thus, given the impartiality of the distance measure, equal treatment will automatically minimize not only expected unfairness but also maximal possible unfairness, if that measure satisfies the Structure Condition.[14] There is

[14] This means, in particular, that equal treatment minimizes maximal possible unfairness in Rankings with Kemeny-Snell distance and in Cake-Divisions with Minkowski distance, but does *not* minimize it in the scholarship example. If $P(\{A\},$

therefore no need for an independent worry about minimax considerations. However, an interesting question is whether there are any plausible interpretations of our model in which the distance measure satisfies Minimax but violates the Structure Condition. On such interpretations, it will be possible to have probability distributions that do not discriminate between individuals and with respect to which **e** does not minimize expected unfairness, even though on all non-discriminative probability distributions **e** will minimize maximal possible unfairness. The issue whether to opt for equal treatment will on such interpretations sometimes depend on whether we adhere to the minimization of expected disvalue or to minimaxing. I don't know, however, of any plausible interpretation of this kind.

7. Extensions

The model we have presented rests on a series of simplifying assumptions. While it always is a good idea to start out with a simple formal framework, this makes our approach unrealistic in several respects. It is therefore natural to consider possible extensions of the modeling. Here are some rather obvious questions to ask:

(i) What if the set of treatments contains *more than one equal treatment*?

In some cases of this kind there might not exist any equal treatment that minimizes expected unfairness or maximal possible unfairness as compared with all other treatments in **T**. What one might hope for, though, is that for every treatment *a* in **T** there is always *some* equal treatment that is at least as satisfactory as *a* in terms of minimization of expected unfairness and/or minimization of maximal possible unfairness. This would mean that treating individuals unequally is never preferable to all forms of equal treatment. More

---

$\{B\}) = P(\{B\}, \{A\}) = ½$, the equal treatment $(\varnothing, \varnothing)$ will be a bad choice for a minimaxer.

precisely, if we just focus on minimization of expected unfairness, what one might hope for is the following:

> *Generalized ExpUnf-minimization:* For every treatment a in $\mathbf{T}$, $\mathbf{T}$ contains *some* equal treatment $\mathbf{e}_a$ such that for every P that does not discriminate between individuals, $\text{ExpUnf}_P(\mathbf{e}_a) \leq \text{ExpUnf}_P(a)$.

It is easy to prove that the condition on the distance measure that is both necessary and sufficient for Generalized ExpUnf-minimization is a generalized version of the Structure Condition:

> *Generalized Structure Condition:* For every $a \in \mathbf{T}$, there exists an equal treatment $\mathbf{e}_a \in \mathbf{T}$ such that for every structure $S \subseteq \mathbf{T}$,
> $$\overline{\mathbf{d}}(\mathbf{e}_a, S) \leq \overline{\mathbf{d}}(a, S).$$[15]

A natural question is whether this Generalized Structure Condition can be derived from some set of more intuitive and basic assumptions about the distance measure.

(ii) What if the available information is indiscriminative between the individuals *within a subgroup* $X \subseteq \mathbf{I}$, but not necessarily outside that subgroup?

Let me introduce some definitions:

> If $X \subseteq \mathbf{I}$, p is an *X-automorphism* iff p is an automorphism that permutes X onto X and for all i in $\mathbf{I}$ that do not belong to X, $p(i) = i$.
>
> A probability distribution P on $\mathbf{T}$ *does not discriminate between the individuals in* $X \subseteq \mathbf{I}$ iff P is invariant under X-automorphisms, i.e, iff for all X-automorphisms p and all $a \in \mathbf{T}$, $P(p(a)) = P(a)$.

[15] For a discussion of Generalized Structure Condition, see footnote 10 above.

If $X \subseteq \mathbf{I}$ and $a, b \in \mathbf{T}$, a and b are *X-structurally identical* iff there is some X-automorphism p such that $p(a) = b$.

a is an X-equal treatment iff for all i, j in X, $a(i) = a(j)$.

If we just focus on the minimization of expected unfairness, we might be interested in the following hypothesis:

Subgroup-Generalized ExpUnf-minimization: For every $X \subseteq \mathbf{I}$ and every $a \in \mathbf{T}$, $\mathbf{T}$ contains *some* X-equal treatment $\mathbf{e}_{a, X}$ such that for every P that does not discriminate between individuals in X, $\text{ExpUnf}_P(\mathbf{e}_{a, X}) \leq \text{ExpUnf}_P(a)$.

This hypothesis provides reasons for equal treatment of those individuals between which our information does not discriminate.
Since X-structural identity is an equivalence relation, just like the ordinary structural identity, $\mathbf{T}$ can be partitioned into *X-structures* – equivalence classes with respect to X-structural identity. The condition on the distance measure that is necessary and sufficient for Subgroup-Generalized ExpUnf-minimization is a further generalization of the Generalized Stucture Condition:

Subgroup-Generalized Structure Condition: For every $X \subseteq \mathbf{I}$ and every $a \in \mathbf{T}$, there exists an X-equal treatment $\mathbf{e}_{a, X} \in \mathbf{T}$ such that for every X-structure $S \subseteq \mathbf{T}$, $\bar{\mathbf{d}}(\mathbf{e}_{a, X}, S) \leq \bar{\mathbf{d}}(a, S)$.[16]

[16] In the case of Cake-Divisions, we may conjecture that $\mathbf{e}_{a,X}$ is obtainable from a as follows: (i) for every i in X, $\mathbf{e}_{a, X}$ assigns to i the average of the values assigned by a to the members of X; while (ii) for every i outside X, $\mathbf{e}_{a, X}(i) = a(i)$. It is less clear, however, how to construct $\mathbf{e}_{a, X}$ in the case of Rankings. The individuals in X should be of course be ranked equally in $\mathbf{e}_{a, X}$. But how should these individuals be ranked vis-à-vis the individuals that do not belong to X? There is no obvious answer to this question.

It is easy to see that this condition entails the Generalized Structure Condition as a special case (for X = **I**), but it would be interesting to know how to derive it from some more basic conditions on a distance measure.

(iii) What if there might be *several (perfectly) fair treatments*, and not just one?

We would then need to work with a different unfairness measure:

> **d**(a, Y) – the degree of unfairness of a on the hypothetical assumption that Y is the set of all fair treatments in **T**.

We take it that **d**(a, Y) is defined only if Y is non-empty and finite. (Finiteness is assumed for the sake of simplicity.) An obvious requirement on this measure is that **d**(a, Y) = 0 iff $a \in Y$. In other words, the degree of unfairness of a is zero if and only if a is one of the (perfectly) fair treatments. In fact, **d**(a, Y) could simply be defined in terms of distance between treatments, as the minimal distance from a to the elements of Y.[17]
We would also need a different measure of information. The new measure would have to be a probability distribution on sets of treatments rather than on individual treatments:

> $P(Y)$ – the probability that Y is the set of all fair treatments in **T**.

For such a probability measure P, the notion of non-discrimination would have to be appropriately re-defined:

> P does not discriminate among the individuals iff for all automorphisms p, $P(p(Y)) = P(Y)$, where $p(Y) = \{b: \exists a \in Y\ p(a) = b\}$.

---

[17] I.e., we could let **d**(a, Y) be min{$b \in Y$: **d**(a, b)}, with **d**(a, b) now interpreted as **d**(a, {b}). If Y were allowed to be infinite, **d**(a, Y) could be instead be identified with the greatest lower bound of {**d**(a, b): $b \in Y$}.

And we would need to correspondingly re-define the notions of expected unfairness and maximal expected unfairness:

$\mathrm{ExpUnf}_P(a)$: $\Sigma_{Y \subseteq \mathbf{T}} P(Y)\, \mathbf{d}(a, Y)$.
$\mathrm{MaxUnf}_P(a) = \max\{\mathbf{d}(a, Y)\colon Y \subseteq \mathbf{T}\ \&\ P(Y) > 0\}$

The obvious question is: What conditions on the re-defined unfairness measure will then guarantee that, in the absence of discriminating information, equal treatment will minimize expected unfairness and/or maximal expected unfairness? Will some condition analogous to our Structure Condition do the job?[18]

These are just some of the follow-up questions that could be raised. But their examination would require another paper.[19]

References

Katzner, L. I. (1973). Presumptions of Reason and Presumptions of Justice. The Journal of Philosophy, 70, 89-100.

---

[18] Here's a natural candidate for such a condition:

If Σ is a family of subsets of **T** such that, for some finite $Y \subseteq \mathbf{T}$, $\Sigma = \{Z \subseteq \mathbf{T}$: for some automorphism p, $p(Y) = Z\}$, then $\bar{\mathbf{d}}(\mathbf{e}, \Sigma) \leq \bar{\mathbf{d}}(a, \Sigma)$ for all $a \in \mathbf{T}$.

If some automorphism permutes *Y* into *Z*, then sets *Y* and *Z* can be said to be structurally identical. The set Σ, which is the set of images of *Y* under automorphisms, can therefore be seen as the *structure* of *Y*. Thus, the suggested condition says the **e** minimizes the average distance to the structure of every finite set of treatments.

[19] This is a significantly revised and streamlined version of Rabinowicz (2010). I am much indebted to the participants of the conference on normative and descriptive models of behaviour at Purdue University, West LaFayette, Indiana, October 2010, for helpful comments and suggestions.

Kemeny, J. G., Snell. J. L. (1962), Preference Rankings. In Kemeny J. G. and J. L. Snell, Mathematical Models in the Social Sciences (pp. 9-23). Cambridge, Mass.: MIT Press, 9-23.

Rabinowicz, W. (2008). Presumption of Equality. In M. L. Jönsson (Ed.), Proceedings of the 2008 Lund-Rutgers Conference (pp. 109-155). Lund: Lund Philosophy Reports.

Rabinowicz, W. (2010). If in Doubt, Treat'em Equally – A Case Study in the Application of Formal Methods to Ethics. In T. Czarnecki, K. Kijania-Placek, O. Poller, J. Wolenski (Eds.), The Analytical Way – Proceedings of the 6th European Congress of Analytic Philosophy pp. 219-243). London: College Publications UK,.

Ullmann-Margalit, E. (1983). On Presumption. The Journal of Philosophy, 80, 143-163.