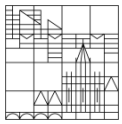
The argument against neutrality about the size of population

David Pomerenke

Supervised by
Stefan Fischer
Jacob Rosenthal

A thesis presented for the degree of
Bachelor of Arts



Department of Philosophy
University of Konstanz
Germany
December 2019

Contents

Introduction

How should we as a society value changes in population size? The question may be crucial when evaluating global warming scenarios. I defend the intuition of neutrality, which answers a part of the question. It states that – other things being equal – it is ethically irrelevant whether or not additional people are added to a population. The argument against neutrality criticizes the intuition to be inconsistent. I present three new objections to the argument: First, economic efficiency needs not be assumed as an ethical principle. Second, the intuition can be interpreted consistently in terms of uncertainty. Third, the intuition can be interpreted and justified in contractarianism. These objections are independent from each other. They are built on controversial philosophical views and do not necessarily disprove the argument against neutrality. Rather, they undermine the authority of the argument by pointing out the weakness of several premises.

I begin by briefly introducing the framework of welfare economics, which this essay argues within. I then present in more detail the intuition of neutrality and the formal argument brought forward against it. The main part is dedicated to the development of three objections to the argument. I conclude with some remarks about the status and plausibility of the different objections.

Chapter 1

Exposition

1.1 Welfare Economics

Welfare economics is the theory how individual well-being should be aggregated to general well-being (or welfare). General well-being drives decisions in the welfare state. The theory is relevant for the execution as well as the design of economic policies. As in democracies the citizens and their representatives take part in the design process, welfare economics is subject to societal discourse in these nations. Within this discourse, citizens and media often do not only claim their own interests. Instead they also refer to ethical principles which are to guide democratic policy decisions. This essay is set within this democratic discourse and aims to defend a supposedly widespread intuition whose consistency has been challenged from the academic side.

The core of welfare economics is the welfare function (see Harsanyi, 1955, p. 309). It is an aggregation function: a function which takes in the individual levels of well-being of several individual persons, and delivers the level of welfare for the whole aggregated population comprising these individual persons. Well-being and welfare (which refers to aggregated well-being) are abstract terms. They are usually interpreted

as a representation derived from a person's preferences about different lives (cf. Crisp, 2017, ch. 4.2). But they can also be interpreted simply as hedonic levels of lifetime pleasure Crisp, 2017, ch. 4.1, which will be sufficient for the purpose of this thesis. Well-being (or utility) of a person p is denoted by $u(p)$; individual persons are denoted by p_i – the subscript is just there to differentiate between different persons. In similar fashion, welfare of a population $P = \{p_1, p_2, \ldots, p_n\}$ is denoted by $u(P)$.

Definition 1: Welfare Function

$$w: \quad \mathbb{R}_+^{|P|} \to \mathbb{R}_+, \qquad \{u(p1), u(p2), \ldots, u(pn)\} \mapsto u(\{p1, p2, \ldots, p_n\})$$

The content of the general welfare function is intentionally unspecified. The function is just a vehicle for discussion within welfare economics. Several specific welfare functions have been proposed and we will deal with two of them in later sections. For example, the classical utilitarian welfare function states that welfare is simply the sum $u(p_1) + u(p_2) + \ldots + u(p_n)$ of all individual well-being.

I introduce welfare functions because they are precise formalizations of competing ethical beliefs. In sections 2.1 and 2.2, I will make use of them in order to demonstrate that when we assume certain ethical intuitions, the argument against neutrality does not hold. I will present two widespread competing ethical belief systems – average utilitarianism, and the difference principle – and try to refute the argument against neutrality from each of these views. The idea is that many people will adhere to one of these principles so that they can agree with at least one of the refutations. (Section 2.3 is of a different kind because, rather than to specific welfare functions, it relates to their justification.)

Welfare economics are blind in a certain respect, and so will be this discussion: They are consequentialist. This means that they only evaluate actions by their outcome and in this context specifically by their impact on general welfare or goodness. Other elements of ethical eval-

uation, such as the procedural requirements of justice, will have to be considered separately (cf. Broome, 2005, p. 401; Broome, 2012, p. 99f). These separate considerations will often require consequentialist considerations as part of their theoretical foundation, so this discussion may be indirectly relevant for them.

1.2 The intuition of neutrality

The intuition of neutrality is assumed to be a widespread ethical intuition among humans (Broome, 2012, p. 176f). The content of the intuition is called the principle of equal existence (Broome, 2004, p. 146), but usually (and also in this thesis) the term "intuition" is also used to refer to the content of the intuition.

The content of the intuition is defined as follows: Let us assume two hypothetical scenarios A and B. The same people exist in both scenarios, except that in scenario B there are some additional people which do not exist in scenario A. The intuition says: Which one of the scenarios is better depends entirely on the well-being of the people who exist in both scenarios, and not at all on the additional people who only exist in B – as long as all the additional people in B have a well-being within a certain neutral range. More specifically, as long as the additional people in B are within the neutral range, scenario A is better in terms of welfare if the people who exist in both populations have a higher welfare in scenario A, and scenario B is better in terms of welfare if the people who exist in both populations have a higher welfare in scenario B.

We can formalize the scenarios as different welfare distributions represented by the welfare functions u_A and u_B. Let P_0 be a population of people who exist in both scenarios but need not have the same levels of well-being in both scenarios. Let P_+ be the population of people who exist only in scenario B. Let $[u_1, u_2]$ be the neutral range of well-being for

added people.

Definition 2: The intuition of neutrality

$$\exists u_1, u_2 :$$
$$(\forall x \in P_+ : u_B(x) \in [u_1, u_2]) \rightarrow$$
$$(u_B(P_0) > u_A(P_0) \rightarrow u_B(P_0 \cup P_+) > u_A(P_0)) \wedge$$
$$(u_B(P_0) < u_A(P_0) \rightarrow u_B(P_0 \cup P_+) < u_A(P_0))$$

The formalization is to be interpreted in the following way: It does not matter in terms of welfare whether there exists an additional person in the population who lives at a moderate level of well-being. There are several moderate levels of well-beings, which form a range between a low moderate level of well-being u_1 and a high moderate level of well-being u_2. If however the additional person is at a very low level of well-being – below u_1 – then the person might matter for the calculation of general well-being. (Arguably, the welfare would decrease because of the added person; though this is not specified by the intuition.) Similarly, if the additional person is at a very high level of well-being – above u_2 – then the person might matter for the calculation of general well-being. (Arguably, the welfare would increase because of the added person.)

There is a variation of the intuition of neutrality where the neutral range has no upper limit, i. e., $u_2 = \infty$ (Broome, 2012, p. 113). This may be a better representation of common belief, and I will come back to it in . Whether the upper limit of the range is finite or infinite is of minor concern for this thesis; it is more important to note that there is *some* neutral range.

some section

If the range is sufficiently large, this might simplify welfare calculations, as the following examples demonstrate:

- An exemplary application of the intuition is the evaluation of road safety (Broome, 2004, p. 144f). In this context, the deaths of people dying in accidents must be weighed against the costs of prevent-

ing them. Whilst this is an ethically difficult problem on its own, one important long-term effect is usually left aside: The well-being of the expected potential offspring of the potentially dying person is completely neglected. One possible justification is the intuition of neutrality: According to the intuition, if we can expect the offspring to live within the neutral range of well-being, it is neither positive nor negative whether they exist or not.

- A second example is the evaluation of different scenarios of global warming (Broome, 2012, p. 170). Global warming is likely to kill many people and thereby to prevent their offspring from existing. On the other hand, global warming may increase poverty, which is associated with higher birth rates. Thanks to the intuition of neutrality we can simply leave both of these effects aside in many of our evaluations – which comes handy as predictions in these domains attend to an enormous amount of uncertainty. Broome, 2012, p. 120ff. sees massive problems if the intuition of neutrality cannot be assumed to apply.

It is important to understand that the intuition of neutrality does not imply neutrality about the consequences on the existing population which are caused by the additional population. These consequences may be negative or positive, leading to contrary political reactions such as China's restrictive one-child-policy and Europe's reproduction-promoting policy (Broome, 2012, p. 169). The consequences on the existing population may well determine whether additional people are good or not. Only the well-being of the additional people themselves does not do so according to the intuition of neutrality.

The question whether the intuition of neutrality is in fact a widespread intuition among humans appears not to have been investigated. It is not necessary for the argument against neutrality to assume such an empirical fact. Neither is it necessary for the refutation of this argument

to assume so. If however this refutation were successful and the integrity of the intuition thus restored, then it would be desirable to investigate the empirical prevalence of the intuition.

1.3 The argument against neutrality

Theorem: The argument against neutrality

The intuition of neutrality (Def. 2) is incorrect.

The argument against neutrality (Broome, 2012, p. 177f, where the graphical figure below is also copied from) concludes that the intuition of neutrality is inconsistent. The argument is a version of the mere addition paradox (Broome, 2004, p. 148) and a modification of the adoption problem (Broome, 2004, p. 161). The argument against neutrality is a reduction to the absurd: It assumes that the intuition of neutrality applies, deduces a contradiction, and thus concludes that the intuition is incorrect.

Premise 1: Intuition of neutrality (P1)

The intuition of neutrality is right (see Definition 2).

The deduction of the contradiction is based on the following counterexample:

Counterexample: The situation in the argument against neutrality

(A1) Let A, B, C be scenarios with corresponding distributions of well-being u_A, u_B, u_C and
populations $P_A = P_0, P_B = P_C = P_0 \cup P+$, and
some person $p \in P_0$
such that $u_A(P_0 \setminus \{p\}) = u_B(P_0 \setminus \{p\}) = u_C(P_0 \setminus \{p\})$.

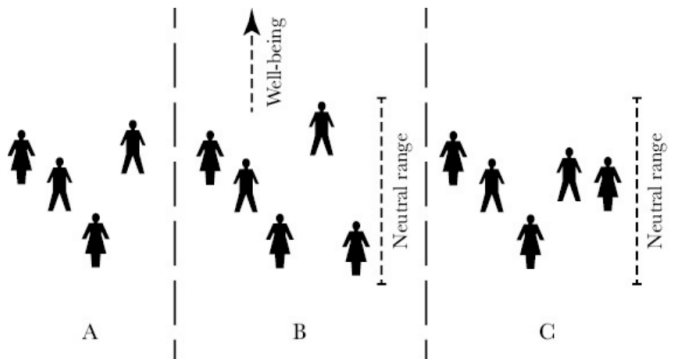
(A2) Let $u_B(p) > u_A(p)$.

(A3) Let $u_C(p) < u_A(p)$.

(A4) Let $P_+ = \{q\}$ with $u_B(q), u_C(q) \in [u_1, u_2]$

(A5) Let $u_B(p) + u_B(q) < u_C(p) + u_C(q)$.

(A6) Let $g_B(P_0 \cup P_+) > g_C(P_0 \cup P+)$.



There are three scenarios A, B and C. They share the same population, except that one additional person exists in both B and C. In both B and C the additional person has a level of well-being within the neutral range. The argument is structured into two major steps:

First, scenario A is being compared to scenario B and to scenario C. The additional person can be neglected in this step because the person is within the neutral range. There is one person who is a little bit better off in scenario B than in scenario A. As all other persons have exactly the

same level of well-being, it is reasonable that there is a higher welfare in scenario B than in scenario A. Contrarily, there is one person who is a little bit worse off in scenario C than in scenario A. As all other persons have exactly the same level of well-being, it is reasonable that there is a higher welfare in scenario A than in scenario C. As a consequence of these two observations, scenario B has a higher welfare than scenario C. Technically, this conclusion requires transitivity of the betterness relation.

Premise 2: Transitivity of betterness (P2)

$$u_X(P_X) > u_Y(P_Y) \wedge u_Y(P_Y) > u_Z(P_Z)$$
$$\rightarrow u_X(P_X) > u_Z(P_Z)$$

Second, scenario B is compared directly to scenario C. Both scenarios comprise the same people, so there is no additional person in either scenario who could be neglected. The person who is not present in scenario A and has therefore been neglected above is much better off in scenario C than in scenario B. This big difference clearly outweighs the difference of the other person's well-being in favour of scenario B. As there is moreover a higher equality in scenario C, scenario C obviously has a higher welfare than scenario B. This is in contradiction to the result of step one, so the counter-example refutes the intuition of neutrality, which has been its core assumption.

Whilst the argument above is intuitively plausible, it has two other important premises (Broome, 2012, p. 177f): First, if in two scenarios all persons have the same level of well-being except for one person who is better off in the second scenario, then the welfare in the second scenario is higher than in the first. Technically, the second scenario Pareto dominates the first (Osborne, 1997).

Premise 3: Pareto domination (P3)

1. $\exists p \in P$:

$$u_X(a) > u_Y(p) \wedge$$
$$\forall q \in P \setminus \{p\} : u_X(q) = u_Y(q))$$
$$\rightarrow u_X(P) > u_Y(P)$$

2. $\exists p \in P$:

$$u_X(a) < u_Y(p) \wedge$$
$$\forall q \in P \setminus \{p\} : u_X(q) = u_Y(q))$$
$$\rightarrow u_X(P) < u_Y(P)$$

Second, if in two scenarios with the same population the sum of individual well-being is higher in the second scenario, and at the same time the inequality of the distribution of well-being is lower in the second scenario, then the second scenario is better in terms of welfare than the first. I call this the fair aggregation principle.

Premise 4: Fair aggregation principle (P4)

$$\sum_{p \in P} u_X(p) > \sum_{p \in P} u_Y(p) \wedge$$
$$g_X(P) < g_Y(P)$$
$$\rightarrow u_X(P) > u_Y(P)$$

with suitable inequality function g (see below).

There are various ways to measure inequality, and the details need not concern us here. An excellent survey of one-dimensional inequality measures – as applied in welfare economics – is given in Sen & Foster, 1997. The most prominent inequality measure is probably the Gini coefficient (see Ceriani & Verme, 2012). Both of these premises appear to be very plausible, and they are dubbed "hard-to-doubt assumptions" in Broome, 2012, p. 176.

The following proof concisely sums up the argument presented above.

It makes use of the technical premises (P...), the assumptions that make up the setting of the counter-example (A...), and the prior conclusions (C...). We can infer from the contradiction in (C8) that at least one of the premises and assumptions must be false. The assumptions merely describe the setting of the scenarios as depicted in the figure above. They are simply the assumptions making up the counter-example and there is no reason to doubt them within this proof. Moreover, premises (P2), (P3) and (P4) appear to be very plausible. As a consequence, the intuition of neutrality must be the false premise.

Proof 1: The argument against neutrality

(C1)	$(P3) \wedge (A1) \wedge (A2) \Rightarrow \ u_B(P_0) > u_A(P_0)$
(C2)	$(P3) \wedge (A1) \wedge (A3) \Rightarrow \ u_C(P_0) < u_A(P_0)$
(C3)	$(C1) \wedge (P1) \wedge (A4) \Rightarrow \ u_B(P_0 \cup P_+) > u_A(P_0)$
(C4)	$(C2) \wedge (P1) \wedge (A4) \Rightarrow \ u_C(P_0 \cup P_+) < u_A(P_0)$
(C5)	$(C3) \wedge (C4) \wedge (P2) \Rightarrow \ u_B(P_0 \cup P_+) > u_C(P_0 \cup P_+)$
(C6)	$(A1) \wedge (A5) \Rightarrow \sum_{x \in P_0} u_B(x) < \sum_{x \in P_0} u_C(x)$
(C7)	$(C6) \wedge (P4) \wedge (A6) \Rightarrow \ u_B(P_0 \cup P_+) < u_C(P_0 \cup P_+)$
(C8)	$(C4) \Leftrightarrow \neg(C7)$
(C9)	$(C8) \wedge (A1 - A6) \wedge (P2) \wedge (P3) \wedge (P4) \Rightarrow \neg(P1)$

There are two major implications if this argument holds and the intuition of neutrality is inconsistent (cf. Broome, 2005, p. 411:

1. We as a society would have to develop a different, consistent principle to replace the intuition. We do not even currently know whether population changes should be evaluated as positive or as negative, just that they cannot simply be evaluated as neutral. The

finding of a new principle with wide acceptance would certainly present a major societal task and require many years of discourse.

2. As soon as we had found a suitable principle, we would need to gain better knowledge of which actions lead to which consequences with respect to population changes. Only then would we probably be able to apply a principle which is not based on neutrality. This requires new scientific analysis and simulation because such predictions have often been omitted in the past (Broome, 2005, p. 402; Broome, 2012, p. 115f).

Broome, 2004 develops five possible responses to the argument against neutrality (see the overview on p. 149). Accordingly, one of the following alternative propositions could be embraced:

(a) intransitivity of the betterness relation

(b) conditional goodness

(c) relative goodness

(d) indeterminacy or vagueness of the betterness relation

(e) a single neutral level

what this mean

what this mean

how this different from uncertainty

The transitivity of the betterness relation (P2) is plausibly defended in (a) – see Broome, 2004, p. 151f. (P3) and (P4) have not been discussed so far. This is what I will do in section 2.1. Section 2.2 will be very similar to what Broome, 2004 develops with regard to proposition (d), but it will also be compatible with proposition (e). I will pursue a somewhat related approach to (b) and (c), focused more on justification, in section 2.3.

aha echt?

Chapter 2

Critical analysis

2.1 Aggregation and justice

I will start by delivering some general criticism on Pareto domination and aggregation and then continue to examine their relation to justice-oriented welfare functions, specifically the Maximin and Leximin rule.

When we say that a scenario Pareto dominates another scenario, we mean that at least one person is better off in this scenario than in the other while all other persons are at an equal level of well-being. The Pareto principle I have formulated as (P2) says that in such cases the first scenario has a higher welfare than the other one. This principle, as well as the extending requirement of Pareto efficiency (cf. Osborne, 1997), find their due place in economics where the objective is the efficient allocation of scarce resources (Samuelson & Nordhaus, 2010, p. 4; Lange, 2019). However I doubt that they are suitable as ethical principles. Pareto efficiency has been criticized because the liberal paradox suggests that it may be incompatible with procedural elements of liberalism (see Sen & Foster, 1997). But I believe that there is a more general problem with Pareto efficiency and even with Pareto domination: Consider a large population with one person whose well-being is much

higher than the well-being of the others. Is it ethically desirable – is there a higher welfare – if the welll-being of this person is increased even more, while the well-being of the other persons remains the same? This can be intuitively doubted, and below I show some mildly convincing reasoning in favour of this doubt.

A similar criticism applies to what I have called the fair aggregation principle. The fair aggregation principle is a combination of what can be called the simple aggregation principle – that general welfare is the simple sum of all individual well-being – with the additional requirement that distributions need be more equal to have a higher welfare. The principle is non-exhaustive: it does not tell us anything about populations with a higher sum of well-being and a lower equality, and it does not tell us anything about populations with a lower sum of well-being and a higher equality. But that is not a problem, since such populations do not play a relevant role in the counter-example to the intuition of neutrality.

The problem with the requirement of equality is that, analytically, equality is a global criterion, which means that it somehow takes into account the well-being of every single person. This implies that a small decrease ϵ in well-being of the person who already is worst off can always be compensated by some large increase of equality within the rest of the population. This follows because otherwise the well-being of the worst-off person would completely determine the equality – which is intuitively plausible, but not incorporated in the conception of inequality measures.

Now imagine three scenarios, all with the same people: In scenario X there is some utility distribution with lots of inequality. The person who is worst off in scenario X is called p. In scenario Y, the well-being of the worst-off person from scenario X is decreased by some very small amount ϵ. Due to the globality of inequality, this can be compensated in terms of equality by improving the equality within all the other persons to a more or less drastical amount. Let us assume that such compensa-

tion has taken place, so that the overall equality in scenario Y is higher than in scenario X. With the usual inequality functions this will be possible without decreasing the sum of well-being (cf. Ceriani & Verme, 2012). Let us further assume that in a third scenario Z all people are at the same level of well-being as the people in scenario Y, plus $\frac{\epsilon}{2}$. The general equality has not decreased in Z in comparison to Y. (Depending on the inequality function, it may even have increased, because the relative differences between the least well-off and the most well-off have decreased.) However the sum of well-being is increased in Z in comparison to X because the well-being of many persons has been increased by $\frac{\epsilon}{2}$ while the well-being of only one person has been decreased by $\frac{\epsilon}{2}$. As a consequence, both the sum of individual well-being and the equality are better in Z than in X, so according to the fair aggregation principle there is a higher welfare in Z than in X. At the same time, the worst-off person in X is even worse off in Z. This seems intuitively implausible and I will now present a theory which explains this implausibility.

STRUCTURE!! 1. pareto stuff, 2. fair aggregation

For this objection I will use as a specific welfare function the difference principle. The difference principle is a concept which is inferred from an analysis of justice. Its justification as the second principle of justice is given and extensively discussed in (Rawls, 2005, pp. 3-183). Rivalling average utilitarianism, the difference principle is probably the most prominent and most widely accepted welfare function. In its core formula, the difference principle states that differences from socioeconomic equality are only permitted if they are to the benefit of the least advantaged (Rawls, 2005, p. 302). This implies that society should aim to optimize the status of the least advantaged. The difference principle is therefore usually represented as a welfare function where general welfare is determined only by the well-being of the group with the lowest level of well-being. (Such representation commits a major error in ignoring the difference between primary goods and well-being as I discuss in Pomerenke, 2017, p. 12f. – But this does not bear upon the reasoning

here, which is based solely on Pareto comparisons.) Whilst the difference principle refers to the least advantaged group – which makes sense in application – there is no mistake in referring to the least advantaged person for the sake of theory (cf. (Rawls, 2005, p. 98)). Because of its resemblance to the decision-theoretic rule of minimum maximization, this formulation of the principle has also been called the Maximin rule. (Although this labelling has been rightly criticized in Rawls & Kelly, 2001, p. 43..)

Definition 3: Difference principle / "Maximin"

$w(P) = \min_{p \in P} u(p)$

According to the difference principle in its Maximin version, both Pareto domination and fair aggregation are false: Imagine that one person who is not the worst-off in either scenario is better off in the first scenario than in the second while all other persons are equally well off. Then Pareto domination requires that the first scenario has a higher welfare. The Maximin rule, however, states that both scenarios have the same welfare because the well-being of the worst-off person has not changed. And we have seen above that as a consequence of fair aggregation a scenario may be evaluated as having a higher welfare even if the worst-off person is even worse off – in strict contradiction to the difference principle.

But the difference principle in its Maximin formulation has been designed as a simplification with the practical idea in mind that there will seldom or never be a comparison in which the least advantaged will have the exactly same level of well-being in both scenarios. Yet for the theoretical case of a such comparison a more elaborate rule than the Maximin rule has been developed (cf. Rawls, 2005, p. 83): It says that in the case that the least advantaged are at the same level in both scenarios, the second-least advantaged must be regarded. And if the

second-least advantaged are also at the same level, then the third-least advantaged must be regarded, and so on. Because it resembles a lexicographical sorting algorithm, the extended rule is called the Leximin rule. It is most clearly formulated as a recursive selection function which outputs the better population of two populations whose members are sorted in ascending order according to their well-being:

Definition 4: Difference principle / "Leximin" selection function

The best population of two populations $S = s_1, ..., s_n$ and $T = t_1, ..., t_n$ which are sorted in ascending by well-being, i. e.,

- $u(s_1) \leq ... \leq u(s_n)$,
- $u(t_1) \leq ... \leq u(t_n)$,

the set of the best population(s) is given by
lexiMin$(S = s_1...s_n, T = t_1...t_n) =$

$$\begin{cases} \{S, T\} & \text{for } S = T = \varnothing \\ \{S\} & \text{for } u(s_1) > u(t_1) \\ \{T\} & \text{for } u(s_1) < u(t_1) \\ \text{lexiMin}\left((s_2, ..., s_n), (t_2, ..., t_n)\right) & \text{for } u(s_1) = u(t_1) \end{cases}$$

We can easily observe that – unlike the Maximin rule – the Leximin rule is compatible with Pareto domination: If all persons are equal in two scenarios except one who is better off in the second scenario, then the Leximin algorithm will recursively call another instance of the Leximin algorithm (where the worst-off from the outer instance will be disregarded), until an instance is called where the two persons in questions are the worst-off persons in their respective scenarios. This process automatically ensures Pareto domination. So at a second glance at the difference principle, it does not contradict but indeed rather support

Pareto domination. This is in favour of the argument against neutrality.

The same, however, cannot be said about the relation of the difference principle to the principle of fair aggregation. We have seen above that fair aggregation in some cases evaluates distributions as being better than a second distribution even though the worst-off person is better off in the second distribution. In such a case, the Leximin algorithm would stop in the first iteration, with a result equivalent to the result of the Maximin rule. The algorithm would not regard the improved well-being of all the other persons, because not only the Maximin rule but also the Leximin rule deem all general improvements irrelevant if they are to the disadvantage of the least advantaged. So for one major welfare function the "hard-to-doubt" premise of fair aggregation (P4) is false and the argument against neutrality cannot succeed.

At the beginning of this section, two intuitive objections to the Pareto principle and the fair aggregation principle have been raised. The objection to the Pareto principle appeared to be supported by assuming the difference principle as a welfare function; however it turned out that the difference principle is only contradictory to the Pareto principle in its Maximin formulation, not in the more general and theoretically preferable formulation as the Leximin rule. The objection to the fair aggregation principle, however, was supported by both the Maximin and the Leximin formulation of the difference principle. As the fair aggregation principle is a necessary premise for the argument against neutrality, the argument therefore fails when the difference principle is assumed as a welfare function.

2.2 Utilitarianism and uncertainty

Whilst section 2.1 has dealt with the implications of assuming the difference principle as a welfare function, this section deals with a second

popular welfare function, that is average utilitarianism.

I will start by explaining how the argument against neutrality requires the neutral range to be a proper range rather than a single level. Afterwards, I will try to make plausible why (assuming average utilitarianism) we should rather assume a single neutral level in theory and elucidate how, taking uncertainty into account, this single neutral level may approach a proper neutral range in practice.

So far, the formalization of the intuition of neutrality involves a neutral range $[u_1, u_2]$ without specifying u_1 or u_2. As per Definition 2, the neutral range could in fact just be a single number with $u_1 = u_2$; but the interpretation of the intuition of neutrality tells us that this range is in fact supposed to be a proper range and rather large.

The argument against neutrality, however, could be misunderstood as an argument against any kind of neutral range. The superficial reader – understanding that the argument denies the possibility of a neutral range – may suppose that it denies the possibility of *any* neutral range. (And this may cause desperation as in .) I want to ward off this potential misunderstanding: As demonstrated below, the argument against neutrality only denies the possibility of a *proper* real range, that is, it denies that the intuition of neutrality holds for $u_1 \neq u_2$.

broome 2012 somewhere

A careful analysis of the argument against neutrality yields that it interprets the intuition of neutrality in a way that does not permit that the neutral range is just a single level of well-being: In order to neutralize and counter the positive difference in well-being for person p between scenarios B and C, the difference in well-being between scenarios B and C for person q must be negative. So the neutral range must allow for such a difference, because the well-being of both p and q is to be within the neutral range:

Corollary 1: Proper neutral range

(C10)	$(A2) \wedge (A3) \Rightarrow u_B(p) > u_C(p)$
(C11)	$(C10) \wedge (A5) \Rightarrow u_B(q) < u_C(q)$
(C12)	$(C11) \wedge (A4) \Rightarrow u_1 < u_2$

Formally (C12), as an implication of the argument against neutrality, is a substantive specification of the intuition of neutrality. Contentwise (C12) is completely in line with the idea behind the intuition of neutrality (cf. Broome, 2004, p. 146): Added lives are neutral except if they are at a very low or very high level of well-being (Broome, 2012, p. 172), so the neutral range is not only a proper range but also a rather big range. The crucial message from the Corollary is that the argument against neutrality has a hidden premise, which has not been made explicit so far: $u_1 < u_2$.

Even if the intuition of neutrality in this form empirically holds as a widespread intuition, it is theoretically problematic:

- One of its implications is for example that we cannot say that a scenario with many added people at the highest well-being within the neutral range is better than a scenario with many added people at the lowest level of well-being. This implication — that well-being within the neutral range is incomparable — is at least controversial.

And there are other pressing theoretical questions:

- What values should u_1 and u_2 assume? Imagine someone proposed as a specification that u_1 should be, say, at the level of well-being of the person at the top of the lowest 10% of the population in terms of well-being. How should we respond?
  - How should we know whether that is correct?

- What kind of arguments would we have to employ in order to plead for a higher or lower value?
- What kind of ethical principle determines the range?

These problems do not arise if we restrict the intuition of neutrality to a single level of neutral well-being:

1. Such a restriction would directly invalidate the argument against neutrality and circumvent the problem of the incomparability of people within the neutral range which I have just touched upon.
2. There exists an established ethical theory which justifies the existence of this level and explains what value it should take.

The theory in (2.) is average utilitarianism and one kind of justification for it is found in Harsanyi, 1955. Average utilitarianism is a highly controversial theory, specifically but not only when it is understood as a complete moral theory rather than only a theory of goodness (cf. Broome, 2012, pp. 50-54; Arrhenius, Ryberg, & Tännsjö, 2017, sec 2.1.1; Rawls, 2005, pp. 167-175, 572f). But it is a popular and consistent ethical theory which is not only able to account for many other ethical intuitions but also to answer our quantitative and justificatory questions regarding the neutral level of well-being. The welfare function of average utilitarianism states that general welfare is the average of all individual well-being:

Definition 5: Average utilitarianism

$$w(P) = \sum_{p \in P} u(p) \cdot |P|^{-1}$$

This implies that in order to be neutral to existing welfare, the welfare of an added population must equal the welfare of the existing population. Not every single added person needs to be at this neutral level, but rather the average of all added persons needs to be at this level.

Definition 6: The neutral range in average utilitarianism

$[u_1, u_2] = u_1 = u_2 = u_0 = w(P)$

So average utilitarianism provides a response to the argument against neutrality by modifying the intuition of neutrality and assuming a neutral level instead of a neutral range. As a result, the intuition is consistent, calculable, and maybe even justified (regarding the justification, cf. Rawls, 2005, pp. 161-175; and also Arrhenius et al., 2017). Average utilitarianism plays (in this case) a revisionist role, a theory of moral error (cf. Mackie, 1990, p. 35): It tells us to slightly adjust our intuition – to sharpen it – so that it is consistent in itself and in its relation to other moral judgments. This is an acceptable, maybe desirable intervention to the beliefs from our intuition.

Furthermore, this theoretical sharpening would not even necessarily change our application of the intuition of neutrality. This is because in practice, uncertainties are attached to all quantities of well-being, specifically the neutral value. When I talk of 'uncertainty' here, then I refer to 'measurement uncertainty' as used in statistics and the quantitative sciences. (The uncertainty in question is quantifiable, so in it falls into the decision-theoretic category of risk and not into the decision-theoretic category of uncertainty.) Measurement uncertainty is a well-developed theory (see, e. g., Runge, 2007). Unlike the approaches of introducing indeterminacy in the forms of incommensurateness or vagueness (cf. alternative (d) in section 1.3) – which are pursued and discarded as a solution to the argument against neutrality in Broome, 2004, pp. 164-183 – uncertainty does not suffer from difficult problems such as greediness (a problem discussed in).

ref to explanation

The neutral value is affected by two kinds of uncertainty:

1. The first kind of uncertainty arises from the definition of the neutral level. Sensitivity analysis (cf. Runge, 2007) of Definition 5 tells us that

the uncertainty of the neutral level is composed of the average uncertainty of the well-being of all existing people.

2. The level of well-being of any actual person that is considered to be at the neutral level or not is also subject to some uncertainty. In both cases, the uncertainty arises from the difficulty to quantify the personal well-being of existing or hypothetical persons (see Harsanyi, 1955, p. 317-319 for discussion).

These uncertainties are not on a theoretical level. On the theoretical level it has been questioned that such quantifications are metaphysically and psychologically possible at all (cf. Harsanyi, 1955, pp. 317-319). – On the practical level, these quantifications are de facto happening, but there is a great level of uncertainty attached to them (Broome, 2004, ch. 9).

We can then accept the theoretical notion of a neutral level while at the same time both maintaining the practical idea of the intuition of neutrality and avoiding the argument against neutrality.

Is doing so just a sophisticated trick? No: The specific nature of the intuition of neutrality had not been analysed before. Rather, it may have been a bit rash to conclude from the rough idea of the intuition of neutrality that it has to be formalized as a proper range.

This section has explained that there is at least no obvious possibility of justifying such a range, and that as a consequence we do not know how to quantify the range. Average utilitarianism presents a possible justification for a neutral level, and together with uncertainty it can justify something like a range. This formal interpretation may be even closer to the empirical intuition of neutrality than the interpretation as a real range is. If it is not, the problem of the incomparability within the neutral range and the problem of the quantifiability of the neutral range present compelling reasons why we should adjust our intuitions.

2.3 Objection 3: Populations act upon their own interests / are selfish

make clear in which case this matters: neither rawls nor avg util – also zb was?

während bei (2) der neutral range auf eine zahl verengt wird, wird er in (3) ausgedehnt, sodass alles neutral ist

The difference principle and average utilitarianism are probably the two most prominent welfare functions. In both of them the argument against neutrality does not hold for different reasons. Yet, although these frameworks are so well received, they both suffer from a justificatory problem. The specific justificatory problem which matters in our context is that the frameworks assume a universal moral domain. This means that they assume that in the first place every person should receive moral consideration. The universal domain is what I will question in this objection and it will lead to another solution of (response to) the argument against neutrality.

So far we have interpreted the intuition of neutrality as a principle which is applied only in a particular instance of comparing the welfare of two scenario: Whenever there are additional persons in one scenario who do not exist in the other scenario, then we can apply the intuition of neutrality. This is reflected in the conclusions in Proof 1. In (C3) and (C4) we have used the intuition of neutrality because there is a different number of persons in scenario A than there is in scenarios B and C. But we have not used the intuition in (C7) because we have been comparing scenarios B and C, and these scenarios have the same number of persons: "B and C contain the very same five people, so in comparing their values all five count as existing people." (Broome, 2012, p. 177; this has been marked as potentially problematic in a talk with Stefan Fischer.)

A simple solution to the argument against against neutrality is to deny

that in such cases all people count as existing people. If we regard person q (the person who exists in scenarios B and C only) as non-existent, then we cannot derive that C is better than B by direct comparison (C7), and the argument fails: (explain ` notation briefly) This requires a revised version of the intuition of neutrality (P1). The only thing which is different from Def. 2 is that the intuition has been extended to comparisons of scenarios where the number of persons is the same: (very unclear! there is no numbers in the formula)

Definition 7: The intuition of neutrality (revised)

$$\exists u_1, u_2 :$$
$$(\forall x \in P_+ : u_B(x) \in [u_1, u_2]) \rightarrow$$
$$(u_B(P_0) > u_A(P_0) \rightarrow u_B(P_0 \cup P_+) > u_A(P_0 \cup P_+)) \wedge$$
$$(u_B(P_0) < u_A(P_0) \rightarrow u_B(P_0 \cup P_+) < u_A(P_0 \cup P_+))$$

But while this solution is compelling so far, it brings with it a formal problem. Thus far, we have not really needed to specify P0 and P+ any more than what is implicit in Def. 2: P0 is the population which exists in both scenarios and P+ is the population which exists only in scenario B. This is no longer implicit in Def. 7: P0 and P+ both exist in both scenarios. (this should probs go before the definition) There is no formal way to distinguish them. P+ are the people who are neutral with respect to general welfare if their well-being lies within the neutral range. And P+ are the same people in B and in C. But P+ could be any persons: P+ could be all persons, no persons, or an arbitrary selection of persons. So as they are not already formally specified we need to specify P+. It is obvious how we specify them:

Definition 8: Additional specification of the intuition of neutrality

P_0 are the existing people and P_+ are the non-existing people.

Unlike all the formal definitions above, this is a material definition,

which is not a problem. The problem is that it is also a relative definition. Which people are existing and which are not depends on the time of evaluation. When we consider whether it is good or not that a baby is born, we arrive at different evaluations before and after the pregnancy of the baby's parent. Before the pregnancy, the baby's well-being has to be ignored because of the intuition of neutrality, but after the pregnancy, the baby's well-being has to be considered. Imagine that we want to know whether it is positive or negative for the general welfare whether the baby suffers from a chronic disease. Then before the pregnancy we will derive that the chronic disease is neutral with respect to general welfare and after the pregnancy we will derive that it would be better for general welfare if the baby does not suffer from a such disease (is better / would have been better).

More formally, the evaluation of welfare depends on what scenario we use as a base scenario based on which we judge which persons are existent and which are not. Such a base scenario may be either of the scenarios which we compare, or a third scenario. In the case of the argument of neutrality, we need to choose scenario A as our base scenario so that we can arrive at the alternative conclusion (C8').

(P5') A is the base scenario,

uA is the distribution of well-being of the base scenario.

So, to be more precise, I revise (C8') and include (P5') as a premise:

(C8') (C7') $\wedge$ (P1) $\wedge$ (P5') $\Rightarrow$ uB(P0 $\cup$ P+) $>$ uC(P0 $\cup$ P+)

A similar approach (how is it different?) to this relativism (in what sense?) is pursued in Broome 2004, pp. 157-162. There it is discarded for two reasons: First, because of the inconsistencies which arise when switching the base scenario (pp. 68-76). Second, because of the difficulty to ethically justify person-relativity or community-relativity (p. 161f). I will now address both issues.

The problem of inconsistency cannot be denied: If ethical evaluations of welfare depend on the choice of the base scenario and if ev-

ery person chooses the person's own situation as the base scenario, inconsistencies will arise. Principally, there are inconsistencies of several kinds. One person could contradict another person from the same population. As we are concerned with population ethics here, where persons will usually somehow consider the whole population for their evaluation, this is not necessarily a problem. A necessary problem is the time-dependence of the evaluation, which is pointed out in Broome, 2004, p. 75: "You choose rightly, but it later turns out you chose wrongly. Indeed, it may turn out that you ought later to undo what you rightly did. Moreover, you might be able to foresee even as you choose A that just this would happen. This is a most implausible sort of incoherence in your activity." This sounds like a problem at first, but in fact it is well acceptable.

(one WEAK possible objection is the idea that) There is no such thing as inconsistency between actions. (This needs SEP backing!!) Such a concept exists of course symbolically, and for example when two persons or one person act out two actions which appear to follow opposite intentions, we might say that the actions are inconsistent. But the concept is very fuzzy and the existing theory of rationality does not provide a criterion for identifying inconsistent actions. What the theory of rationality does provide, is a criterion for identifying inconsistent beliefs: Beliefs are inconsistent if their propositions are contradictory. Without a formal theory of inconsistent actions, philosophers should probably restrict themselves to the analysis of inconsistent beliefs, not actions. (not yet, but why couldn't there be one?) The underlying beliefs are complex: Before the action A, we think that we should do A. And we think that as a causal effect of doing A, we will regret having done A. So we think that we should do A, and that we will regret it afterwards. After the action A, we think that we should not have done A. We also think that we have thought that we should do A. There is no formal contradiction in these beliefs. It may very well be rational to think A at the

moment and to expect that oneself would think the opposite of A under different circumstances. As the enactment of A causes a change of circumstances, the above beliefs may come about, and there is nothing wrong with them. There is also no such thing as a problem of "undoing" an action in philosophical terms. A is done within one set of circumstances and then within another set of circumstances is "undone". The two actions of "doing A" and "undoing A" can be differentiated by the fact that they have taken place within different contexts (one context without the causal effects of A and one context with the causal effects of A). There is no reason to ignore the contexts of the actions and to strip them down to the notions of "doing A" and "undoing A".

Let us examine two examples where we ignore any ethical constraints and just consider the consistency of their beliefs. Imagine a community of two persons who reason whether or not they should improve some genetic condition of their child. They may be concerned about the well-being of the family and think: "Our family is rather against this for religious reasons. Once the child is born, it will clearly prefer that the modification had taken place. Our family will then have to respect the preferences of the child. As a result, our family will think that the modification should have taken place; and if it will not have taken place, our family will probably be very unhappy about it." The parents in their situation of reasoning should then consider whether they want to experience this later unhappiness. But there is no rule of rationality which would require them to somehow align the present and future preferences of the family. After all, the family will consist of different persons after the child is born. As a second example, imagine a direct democracy called Alphaland whose citizens consider – because of their liberal ideal – to invade and annex an autocratic country called Betaland with a population bigger than their own population. They consensually adopt a resolution: "Alphaland wants to annex Betaland. Alphaland expects Betaland to condemn the violence related to the annexation. As the current citizens of Beta-

land will be the majority in Alphaland after the annexation, we expect that Alphaland will officially regret the annexation afterwards. But we expect that most former Betaland citizens will nonetheless want to remain in Alphaland for pragmatic reasons (so there is no reason against the action)." Rationality does not require present Alphaland to consider the interest future Alphaland. After all, future Alphaland is made up of different citizens than present Alphaland. These examples illustrate that there is no requirement for communities to be consistent with their beliefs and expected beliefs over time. (Whether this principle also applies to individual agents is up for discussion, but luckily needs not concern us here.)

hier deutlicher das problem auspointen: communities erscheinen, zb weil sie namen haben, wie menschen. dabei gelten viel weniger strenge regeln zb bzgl konsistenz (vgl auch arrow). ob community denken gut ist, bleibt offen, aber es passiert ja offenbar. evtl wäre der absatz über diesem ein argument gegen solches community denken (will ich nicht FÜR community denken plädieren?)

I have just talked about the consistency of beliefs of communities. But we are here concerned with general ethical belief, which one would assume to be independent from the practical belief of some community. (bessere überleitung) This disparity is the subject of the second objection to relativism: Ethics is nothing related to the welfare of any community, but it is related to the welfare of all – ethics has a universal domain.

- i do _not_ want to say that ethics applies population-wise (as one could think from my examples above). it applies of course to the complete domain of humans, because all humans have similar interests (Stemmer 2000).
- infer intuition of neutrality with a real range from egoistic interests
  - it is unlikely that all or many people would grant rights to nonexisting

- there is no per se effect of happiness on other people's happiness, but in the extremes it is highly likely
  - mitleidsethik
  - psychological problem: people do not evaluate the effect of mitleid as a change of their own welfare, but as a change of global welfare!
  - -> revisionism: when these effects are really properly included in the welfare function, then we can have an ion that they are completely irrelevant

Conclusion

Contractarianism	Average Utilitarianism	Difference Principle	Note	Example	Compatible
✓	✓	×	Utilitarianism with contractarian justification	Harsanyi, 1955	✓
✓	×	✓	Difference principle with contractarian justification	Rawls, 2005	✓
✓	×	×	Contractarianism with other or no welfare function	Stemmer, 2000	✓
×	✓	×	Utilitarianism with other justification	Broome, 2004	✓
×	×	✓	Difference principle with other justification	cf. Pomerenke, 2017	✓
×	×	×	Other moral framework		N.N.

- explain why objection 3 is most important: 1 and 2 rely on rawls and avg util, which are basically based on an analytic explanation of – within our linguistic framework – justice as fairness and ethics as impartiality, resp., whilst 3 is based on the interest of people and thus normative

In summary, Broome commits two errors: In two common frameworks his "hard-to-doubt" strong Pareto assumption is false. In the third framework his implicit assumption that the intuition of neutrality requires proper ranges of neutral values is false. It is also plausible that the assumptions

on their own do not reflect common ethical intuition. The intuition of neutrality might actually be right.

- point out conclusion: ion can maybe be held in utilitarianism (revision: no range but uncertainty), for sure in rawls & similar frameworks, most plausibly in contractarianism (revision: no upper / lower boundaries)
- sum up frameworks, discuss whether they are a complete partitioning of ethical belief
- (Assuming that ethics is a system which should be built up from people's ethical convictions,) we need empirical research on whether the intuition holds
- impact on climate change? we need not care about unborn people and can make more straight calculations

8221 words, 51837 characters

References

Arrhenius, G., Ryberg, J., & Tännsjö, T. (2017). The Repugnant Conclusion. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 ed.). Metaphysics Research Lab, Stanford University. Retrieved 2019-08-30, from https://plato.stanford.edu/archives/spr2017/entries/repugnant-conclusion/

Broome, J. (2004). *Weighing lives*. New York: Oxford University Press.

Broome, J. (2005). Should we value population? *Journal of Political Philosophy*, *13*(4), 399–413.

Broome, J. (2012). *Climate matters: Ethics in a warming world (Norton global ethics series)*. WW Norton and Company.

Ceriani, L., & Verme, P. (2012). The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *The Journal of Economic Inequality*, *10*(3), 421–443.

Crisp, R. (2017). Well-Being. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2017 ed.). Metaphysics Research Lab, Stanford University. Retrieved 2019-12-25, from https://plato.stanford.edu/archives/fall2017/entries/well-being/

Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of political economy*, *63*(4), 309–321.

Lange, L. (2019, August). *Definition of economics*.

Mackie, J. (1990). *Ethics: Inventing right and wrong*. Penguin UK.

Osborne, M. J. (1997). *Pareto efficiency.* Retrieved 2019-08-27, from https://www.economics.utoronto.ca/osborne/2x3/tutorial/PE.HTM

Pomerenke, D. (2017). *Nach welchen Prinzipien sollte der Staat die Verteilung von Gütern gestalten? Eine systematische Darstellung der Diskussion zwischen John Rawls und John Harsanyi* (Hausarbeit). Konstanz. Retrieved from https://archive.org/details/rawls_vs_harsanyi

Rawls, J. (2005). *A theory of justice* (Original ed ed.). Cambridge, Massachusetts: Belknap Press.

Rawls, J., & Kelly, E. (2001). *Justice as fairness: a restatement*. Cambridge, Mass: Harvard University Press.

Runge, B.-U. (2007). Messunsicherheitsanalyse. In *Physikalisches Anfängerpraktikum. Skript.* Universität Kontanz. Retrieved from https://ap.physik.uni-konstanz.de/index.php?Itemid=19

Samuelson, P. A., & Nordhaus, W. D. (2010). *Economics* (19th ed ed.). Boston: McGraw-Hill Irwin.

Sen, A., & Foster, J. E. (1997). *On Economic Inequality*. Clarendon Press.

Stemmer, P. (2000). *Handeln zugunsten anderer: eine moralphilosophische Untersuchung*. Berlin ; New York: W. de Gruyter.