Variable critical level utilitarianism

Abstract

Variable critical level utilitarianism is an extension of critical level theories in population ethics. It assumes that people are free to choose their own critical levels that are included in the welfare function. These critical levels can also differ between situations and can even depend on the choice sets of possible situations. Traditional population ethical theories, such as total and average utilitarianism and person affecting views, are limiting cases of variable critical level utilitarianism. Those traditional theories face counterintuitive implications. The flexibility of variable critical level utilitarianism allows to avoid those population ethical problems. Nevertheless, variable critical level utilitarianism has a game theoretic dynamic inconsistency, that is explained using examples of livestock farming and climate change.

Introduction

Population ethics is an area in moral philosophy that studies which choices are the best when the choices determine not only the welfare of people but also the size of populations and the existence or non-existence of individuals. Think about problems in bio-ethics, such as fertility treatments or prenatal genetic screening, or problems in animal ethics, such as wildlife population control or domestic animal breeding. Population ethics also has important implications in economic cost-benefit analysis and prioritization research. Should we prioritize helping current people or preventing harm to future people? How to measure the cost of human extinction from a catastrophic risk? How to measure harms to future people, when preventive measures imply that other people will exist in the future? Should future welfare be discounted?

In a consequentialist-axiological approach to population ethics, we are looking for a welfare function, a mathematical tool to select the best situation. The best choice is the one that maximizes this welfare function. When this welfare function is an aggregate of the utilities of all individuals, we get a utilitarian theory. The utility of an individual is a function of everything that the individual values or prefers, such as happiness or well-being. (For simplicity, I use utility and happiness as synonyms in this article).

Since Derek Parfit (1984) brought attention to paradoxes and repugnant conclusions of population ethical theories, many other researchers started looking for refined consequentialist-utilitarian theories that avoid the many counterintuitive implications (Arrhenius, 2000; Blackorby, Bossert & Donaldson, 2005; Mulgan, 2006). However, as in social choice theory (Arrow, 1950) and bargaining theory (Myerson & Satterthwaite, 1983), mathematical impossibility theorems in population ethics were derived (Arrhenius, 2000b; Blackorby, Bossert & Donaldson, 2003). It is impossible to construct an axiological social welfare function that ranks all social states and is maximized at the social state that is most preferred according to some very basic moral intuitions. Some intuitively appealing conditions for a population ethical theory are mutually inconsistent. We have for example a moral intuition that increasing the happiness of existing people by adding extra happy people, cannot make a situation worse, and the intuition that adding an extra person with negative welfare cannot make a situation better, all else equal. However, it can be proven mathematically that always at least one of such basic intuitions has to be violated. So the best population ethical theory is the one that only violates the weakest moral intuitions, the softest bullets to bite.

In this article I present a new candidate of such a best population ethical theory. It resembles critical level utilitarianism (Blackorby, Bossert & Donaldson, 1995, 1997), with one simple change: people who exist are completely free to determine their own critical levels, which can differ from situation to situation. Hence, the critical levels are not constants, but they are variable. I will first describe this variable critical level utilitarianism. Second, I will demonstrate how traditional population ethical theories in the literature are limiting cases of variable critical level utilitarianism, and how these traditional theories face very counterintuitive problems. Finally, I will use game theory considerations to show how variable critical level utilitarianism can avoid those most counterintuitive problems, and also how a problem like dynamic inconsistency (subgame imperfect equilibria) can arise. Nevertheless, this dynamic inconsistency vice of variable critical level utilitarianism can also be interpreted as a virtue, because it can help explain some intuitive moral judgments about e.g. meat consumption, slavery or climate change.

Full variable critical level utilitarianism

Variable critical level utilitarianism (VCLU) starts with relative utilities of individuals

$$u_{rel}(i, s, \mathcal{C}) = u(i, s) - c(i, s, \mathcal{C}).$$

The utility $u(i, s)$ of individual i who exists in situation s measures the preference of that individual for situation s. A person with zero utility (e.g. having neither positive nor negative experiences) is indifferent with non-existence. The critical level $c(i, s, \mathcal{C})$ chosen by individual i in situation s, given the choice set $\mathcal{C}$ of all situations that can be chosen (by a social planner or society), reflects a maximum willingness to pay (in utility terms) of individual i to avoid situation s. The latter interpretation will become clear later in the section with the game theory analogy.

With the relative utilities we can construct the social welfare function as the sum of everyone's relative utilities. If a person does not exist in situation s, both its utility and critical level are zero. The social optimum situation s_{opt} is the one that maximizes the social welfare function:

$$s_{opt}(\mathcal{C}) = \arg\max_{s \in \mathcal{C}} \sum_{i \in P(s)} \big(u(i, s) - c(i, s, \mathcal{C})\big),$$

where the sum runs over all individuals who belong to the population $P(s)$ of situation s. As the maximization runs over all situations that are elements of the choice set $\mathcal{C}$, the optimum situation depends on the choice set.

A full variable critical level theory allows the individuals to be free to set their own critical levels. This maximally respects autonomy of individuals. However, three constraints can be imposed.

First, if a person does not exist in a situation, that non-existing person is not allowed (or able) to choose a critical level for that situation. So we have $c(i, s, \mathcal{C}) = 0$ if $i \notin P(s)$, i.e. if the person is no member of the population $P(s)$ of situation s (the set of individuals existing in that situation).

Second, an individual can only choose a non-negative critical level. So we require $c(i, s, \mathcal{C}) \in \mathbb{R}_{\geq 0}$, for all individuals, situations and choice sets, also if the individual has a positive utility $u(i, s) \geq 0$. This is a rationality constraint: if a person would choose a negative critical level, that person kind of acknowledges that his or her existence can improve the social welfare, even if that person would have a negative utility. Or in other words: an individual should be willing to accept a life with a utility equal to the chosen critical level, and no-one could reasonably accept a life with negative utility.

Third, the total critical level, i.e. the sum of all critical levels set by the individuals in a situation, has an upper bound, given by the maximum over all possible situations of the sum of positive utilities of the people who exist in that situation but do not exist in all possible situations. This restriction will be explained in a later section, but here it suffices to say that this restriction is required to avoid that people choose infinite critical levels.

These three restrictions are the rules of the game of full variable critical level utilitarianism. We can impose further restrictions that make the critical levels less variable. As the critical levels are functions of three arguments, there are three natural restrictions.

Person independent VCLU says that everyone should choose the same critical level, but those levels can be different in different situations and choice sets: for all s and $\mathcal{C}$, the critical levels are $c(i, s, \mathcal{C}) = c(i', s, \mathcal{C}) = c(s, \mathcal{C})$ for all i and i'.

Situation independent VCLU says that a person that exists in several situations, has to choose the same critical level in all those situations: for all i and $\mathcal{C}$, the critical levels are $c(i, s, \mathcal{C}) = c(i, s', \mathcal{C}) = c(i, \mathcal{C})$ for all s and s' that include the person. Note that it is not always clear to say that a person in one situation is the same person that exists in another situation. For example, in one situation you grew up in the environment that you lived in, and you became the person that you are now. In another situation, you were the same baby (with the same genes), but you grew up in a completely different environment, such that you adopted other memories, desires, feelings and so on. Are you the same person as your alter ego in that other situation? This identity can be a matter of degree. Or what about science fiction thought experiments about copying your mind, teleporting your brain, emulating your brain on a computer, changing your neurons one by one,…? To avoid discussions about personal identity, situation independent VCLU imposes that once person i in situation S identifies himself with his 'alter-ego' i in situation s', person i in situation s has to pick $c(i, s', \mathcal{C})$ as is own critical level. A person is free to identify herself with anyone else in any other situation and this identification does not have to be symmetrical.

Choice set independent VCLU says that the chosen critical levels should be independent from the choice set: for all i and s, the critical levels are $c(i, s, \mathcal{C}) = c(i, s, \mathcal{C}') = c(i, s')$ for all $\mathcal{C}$ and $\mathcal{C}'$.

As we will see in the next section, many traditional population ethical theories can be described as person, situation or choice set independent restrictions of VCLU. These restrictions can generate the counterintuitive implications of those theories. And they violate full autonomy of individuals to choose their own critical levels.

Population ethical theories

This section presents an overview of the major traditional population ethical theories, and how they can be considered as special cases of VCLU.

Number-sensitive critical level theories

General expression

Number-sensitive critical level utilitarianism (Blackorby, Bossert & Donaldson, 2002) uses as critical level: $c_{NCLU}(s) = g(N(s))$, where g is a general, non-negative function of the population size (the number N of existing people in situation s).

Variations and limiting cases

If the function g is a constant, independent of population size, we arrive at the original, constant critical level utilitarianism (Blackorby, Bossert & Donaldson, 1995; 1997), where $c(i, s, \mathcal{C}) = c_{CLU}$ is independent from person, situation and choice set. If this critical level is zero, we get total utilitarianism: $c(i, s, \mathcal{C}) = c_{TU} = 0$.

Counterexamples

Number-sensitive critical level theories have counterintuitive implications. In this section I describe the most serious objections, i.e. the toughest bullets to bite if one is willing to accept one of those theories.

If $g(N)$ strictly increases in N for large N, it can be easily shown that the theory implies the:

> *Strong Dominance Addition Problem 1* (Arrhenius 2000, p159). Situation s contains N very happy people with slight happiness inequality. Situation s' contains the same happy people, with higher and equal happiness level, plus an extra person with the same high happiness level as the others.

In situation s', an extra person is added, which results in a higher total and average utility and a perfect equality. Hence, s' is dominating s in many aspects that people deem valuable. But according to number-sensitive critical level utilitarianism, the social value (social welfare function) is higher for s when N is sufficiently large. That means adding the extra person would make things worse. If a theory says that the addition of extra people, which results in a situation that dominates other situations in relevant aspects (e.g. in terms of total and average utility and equality), is strictly worse than the other situations, the theory faces the strong dominance addition problem.[1]

When $g(N)$ is not constant, number-sensitive critical level utilitarianism also faces the:

> *Dependency Problem*: whether adding extra very happy people is good or bad depends on how many other people already exist elsewhere.

For example, it can be the case that adding an extra happy person is good when there already exist N people, but when there happen to be more people (for example when other planets with sentient life forms are discovered), adding an extra happy person can become bad.

If $g(N)$ approaches or fluctuates around a constant for large N, as is also the case in the original critical level and total utilitarian theories, we face sadistic conclusions. If the limit critical level is high we get the:

> *Sadistic Problem 1* (Arrhenius 2000, p63). Situation s contains N people with extreme suffering (very negative utility). Situation s' contains the same N people with extreme happiness, plus M extra people with high positive utilities slightly below the critical level.

If M is large enough, the critical level theory prefers s above s'. This is a sadistic conclusion, because the people in situation s are suffering.

If the limit critical level is low (e.g. zero as in total utilitarianism), we get the:

[1] If the theory says that the new situation is not better, without implying that it is worse, then we have a weak version of the dominance addition problem. In that case, the two situations can be equally good or be incomparable.

> *Sadistic Problem 2*. Situation s contains N people with extreme happiness. Situation s' contains the same N people with extreme misery, plus M extra people with lives barely worth living, i.e. positive utilities slightly above zero (or slightly above the low critical level).

If M is large enough, this critical level theory chooses s' above s, and hence picks the situation where N people are extremely miserable. The huge number of extra people with lives barely worth living can have a huge total utility, but still each individual life is barely worth living. The conclusion that total utilitarianism chooses a situation with many lives barely worth living above a situation with a small number of extremally happy people, is known as the repugnant conclusion (Parfit, 1984). Sadistic conclusion 2 is a stronger version of this repugnant conclusion, which makes total utilitarianism even more counterintuitive.

Average-utility-sensitive critical level theories

General expression

Variable value (or number-dampened) utilitarianism (Hurka, 1983; Ng, 1989) uses as a situation dependent critical level a function of population size $N(s)$ and average utility: $c_{VU}(s) = \frac{(N-f(N))}{N}\frac{\sum_i u(i,s)}{N}$. With these values, the welfare function becomes

$$W_{VU} = \frac{f(N(s))}{N(s)}\sum_i u(i,s).$$

Variations and limiting cases

When $f(N) = N$, we again arrive at total utilitarianism: $c_{TU} = 0$. When $f(N) = 1$, we arrive at average utilitarianism: $c_{AU}(s) = \frac{(N-1)}{N}\frac{\sum_i u(i,s)}{N}$. In this case, the welfare function is simply the average utility of the population.

Counterexamples

If $f(N)$ is strictly increasing in N for large N (as with total utilitarianism), variable value utilitarianism faces the abovementioned sadistic conclusion 2. If $f(N)$ approaches a constant for large N (as with average utilitarianism), we get the

> *Sadistic Problem 3*: Situation s contains N people with extreme happiness and 1 person with extreme suffering. Situation s' contains the same N people with extreme happiness, the same 1 person, but with extreme happiness, plus extra M people with lives close to but slightly lower than extreme happiness.

According to average utilitarianism, if M is large enough, then s would be preferred, even if one person in s suffers a lot. Relieving his suffering, making him absolutely happy, by adding many extra extremely happy people, could still decrease average utility and hence decrease the welfare function. Average utilitarianism also faces the abovementioned dependency problem: whether adding an extra happy person is good or bad depends on the existence of other people elsewhere. If we discover a planet with a huge population of extremely happy people, we should not procreate, because our children will have a lower happiness level and hence lower the average utility in the world.

Personal-utility-sensitive critical level theories

General expression

According to utility-sensitive critical level utilitarianism, the person and situation dependent critical levels are a function of the individual utilities: $c_{SCLU}(i,s) = h(u(i,s))$ with $h(u)$ a positive, decreasing function with horizontal asymptote at zero.

Variations and limiting cases

When $h(u)$ is a step function, we get a sufficientarian critical level theory (resembling the one described by Shields, 2012): $c_{SCLU}(i,s) = c > \hat{c}$ if $u(i,s) \leq \hat{c}$ and $c_{SCLU}(i,s) = 0$ if $u(i,s) > \hat{c}$. This is a two-tier critical level theory with pivotal critical level $\hat{c}$.

Counterexamples

It is easy to show that a sufficientarian theory implies the sadistic conclusion 1 if the pivotal critical level is high, and the sadistic conclusion 2 if the pivotal critical level is low.

Satisficing theories

General expression

In a satisficing utilitarian theory, the person, situation and choice set dependent critical level is the maximum of the actual individual utility in the situation and the sufficiency level c_S (which can be individual, situation and choice set dependent): $c_{SU}(i,s,\mathcal{C}) = \max\{u(i,s), c_S(i,s,\mathcal{C})\}$. If the utility is higher than the sufficient level, the relative utility becomes zero, and hence the person's welfare no longer counts.

Variations and limiting cases

Antifrustrationism (Fehige, 1992) can be restated in a satisficing utilitarian theory, with the critical level equal to the satisficing level equal to a maximum utility in a perfect life with no preference frustration: $c_{AF} = c_S = u_P$. Or, more realistically, the satisficing level can be the maximum utility that a person can get: $c_S(\mathcal{C}) = \max_{i \in P(s), s \in \mathcal{C}} u(i,s)$, with $P(s)$ the population of situation s. Also negative utilitarianism (Walker, 1974) can be reframed in a population ethical version, with satisficing level $c_S = 0$ and hence the critical level $c_{NU}(i,s) = \max\{u(i,s), 0\}$. In this theory, only people with negative utilities count.

Meacham (2012) discussed a theory which he called saturating harm minimizing utilitarianism (SHMU). This theory has some resemblance with satisficing utilitarian (and person-affecting) theories. It can be rephrased in terms of a variable critical level theory as follows. Rank all the individuals in situation s according to decreasing utilities. The i-th individual (with the i-th highest utility level in the population of situation s) can be indicated with $i_\downarrow$. For example, the $1_\downarrow$-th person has the highest utility. If the population as $N(s)$ individuals, the utility $u(i_\downarrow, s) = 0$ if $i_\downarrow > N(s)$ (i.e. if situation s does not contain an individual with such a high rank). The SHMU critical level depends on the choice set and the individual, and is defined as $c_{SHMU}(i_\downarrow, \mathcal{C}) = \max\{0, \max_{s \in \mathcal{C}} u(i_\downarrow, s)\}$. Meacham defines the harm of individual $i_\downarrow$ in situation s as the negative of the relative utility $u(i_\downarrow, s) - c_{SHMU}(i_\downarrow, \mathcal{C})$, and hence the total harm should be minimized. The 'saturating' aspect of the theory is related to the matching procedure, where the $i_\downarrow$-th individual in situation s is matched to the $i_\downarrow$-th individual in situation s'. Note that this does not imply personal identity. For example the happiest person in situation s does not need to be the same person as the happiest person in situation s'.

Counterexamples

If c_S is high, the satisficing theories imply the strong dominance addition problem 1. More generally, the theories face the non-existence problem: it is better not to exist, even if existence means a happy life. For example, when the satisficing level is the maximum possible utility, only the person that can get this level of utility is allowed to exist, and even then her existence does not add anything to the welfare function, so her existence is neutral instead of positive.

If c_S is low, the theories are insensitive to positive utility levels of people. Increasing the happiness of a happy person would make no difference. There is a weak dominance addition problem: a situation with low total and average utility and high inequality is not necessarily worse than a situation where extra happy people are added, total and average happiness is higher and inequality is lower.

Concerning SHMU, following the proof of theorem 1 in Blackorby, Bossert, & Donaldson (2003), this theory is vulnerable to the above Sadistic Problem 3: in some choice sets that include situations s and s', SHMU would prefer s.

Person-affecting theories

General expression

An asymmetrical person-affecting utilitarianism (Narveson, 1973; Heyd 1988; Bykvist, 1998; Arrhenius 2000 p128) uses as person, situation and choice set dependent critical levels $c_{APAU}(i,s,\mathcal{C}) = \max\{u(i,s),0\}$ if $\exists s' \in \mathcal{C}$ with $i \notin P(s')$, i.e. if there exists a possible situation s' in the choice set and individual i is not a member of the population of that situation (i.e. she does not exist in that situation). If there is no such situation s', the individual necessarily exists (i.e. in all possible situations) and she is allowed to freely choose her own critical level.

In this theory, there is an asymmetry: adding an extra person with positive utility does not increase the welfare function, but adding an extra person with negative utility does decrease the welfare function.

Variations and limiting cases

If the critical level of a new person is simply her utility (i.e. not the maximum of her utility and zero), we get a symmetrical person-affecting utilitarianism: $c_{SPAU}(i,s,\mathcal{C}) = u(i,s)$ if $\exists s' \in \mathcal{C}$ with $i \notin P(s')$. In this case, adding an extra person with negative utility does not decrease the welfare function.

Asymmetric person-affecting theories can differ due to different restrictions on the choice set. In comparativist person-affecting utilitarianism (Arrhenius, 2000, p119): the choice set $\mathcal{C}$ contains two elements, i.e. this theory only allows for comparisons between two situations. In necessitarian person-affecting utilitarianism the choice set contains all possible situations that one can choose. A person is contingent if there is at least one possible situation in which that person does not exist. If a person exists in all possible situations, the person exists necessarily, and only these persons count in this person-affecting theory. In a presentist person-affecting theory, the choice set contains all hypothetical situations that are compatible with the present situation (including situations that one cannot choose). It means that $c_{PPAU}(i,s,\mathcal{C}) = u(i,s)$ for all individuals i who do not exist at the present (because for all possible future people, there is always another hypothetical situation where those people do not exist). Finally, there is a soft asymmetrical person-affecting utilitarianism (Dasgupta, 1988, p120), with critical level $c_{SAPAU}(i,s,\mathcal{C}) = \max\{u(i,s)/2,0\}$ if $\exists s' \in \mathcal{C}$ with $i \notin$

$P(s')$. For contingent people with positive utilities, their existence contributes $u(i,s)/2$ to the welfare function.

Counterexamples

Symmetrical person-affecting utilitarianism faces a sadistic conclusion. Situation s contains N happy people. Situation s' contains the same N people, slightly happier, plus M extra people with extreme misery. The welfare function of situation s' is higher in this theory. This sadistic conclusion can be easily avoided with an asymmetrical person-affecting theory. This example explains why we restrict critical levels to positive values.

Soft person-affecting utilitarianism is an intermediate case between asymmetric person-effecting utilitarianism and total utilitarianism and hence faces the abovementioned sadistic conclusion 2 of total utilitarianism.

Comparativist person-affecting utilitarianism faces an intransitivity. Situation s contains a very happy person X and a barely happy person Y. In situation s', person X does not exist, person Y is very happy, and another, barely happy person Z exist. In situation s'', person Y does not exist, person Z is very happy and person X exists and is barely happy. Comparing situations s and s', s' is preferred because person Y is the only person that exists in both situations, and he has a higher happiness level in situation s'. Comparing situations s' and s'', the latter is preferred. But comparing s'' and s, situation s is the best. Hence, there is an intransitive ranking s > s'' > s' > s.

All person-affecting theories face the

> *Strong Dominance Addition Problem 2*. Situation s contains N happy people, M extremely happy people plus a huge number L of extra people with the same happiness as the M-people. Situation s' contains the same N happy people, the same M people who are happier than in situation s, plus a small number Q of extra people, not existing in situation s, with lives barely worth living.

With a person-affecting utilitarian welfare function, situation s' is preferred, even if s has higher total and average utility and more equality. All asymmetrical person-affecting theories have the

> *Extreme Priority Problem* (Arrhenius 2000, p128). Situation s contains N happy people and one person with a life barely worth living (positive but close to zero utility). Situation s' contains the same N equally happy people, the same one person, but with a very small negative utility (close to zero), plus a huge number of M extremely happy people.

According to asymmetrical person-affecting utilitarianism, the welfare of the one person that has a life slightly below zero gets an extremely high priority: situation s is chosen because that person gets a slightly higher, positive utility. This is counterintuitive, because total and average utility in s' is much higher, and the utility levels of the one person is almost the same in the two situations.

Finally, necessitarian and presentist person-affecting theories are insensitive to the utilities of possible people and hence face a

> *Weak Dominance Addition Problem*. Situation s contains N very happy people. Situation s' contains the same N people, slightly happier, plus a huge number of M extra, extremely happy people. Situation s'' contains the same N people, as happy as in situation s', plus the same M people as in s' but with lives barely worth living, plus one extra person with a life barely worth living.

When the three options s, s’ and s’’ are possible, person-affecting utilitarianism chooses either s’ or s’’ and is indifferent between these two, even if situation s’ has a higher average and total utility and more equality. Once option s’’ is chosen, this becomes the only choice that is left possible. So s’’ gets the highest value once it is chosen, even if the same M people could have been much happier in situation s’.

Game-theoretic considerations of variable critical level utilitarianism

Variable critical level utilitarianism can be described as a strategic game. Consider a building with a lot of rooms. Outside the building are the players waiting. Before the players may enter the building, the game master explains the set-up. Only when a player’s name is written on the door of a room, the player is allowed in that room. In each room, a player (who is allowed to enter that room) can expect a room specific payoff[2]: a reward (the player receives an amount of money from the game master) or a punishment (the player has to pay an amount of money to the game master). Before they enter the building, all players are fully informed about the payoffs they and the other players can get in all the rooms in which they are allowed.

For each of the permissible rooms, a player can declare his or her avoidance amount: the willingness to pay to avoid that room. Players cannot declare an avoidance amount for the rooms in which they are not allowed. With these declared avoidance amounts, and the payoffs of the players in each room, the game master computes a net welfare value for each room: the sum of the payoffs of all allowed players in the room minus the sum of the declared avoidance amounts of those players.

The room with the highest net welfare level is selected (in case of a tie, one of the rooms is randomly selected). The game master opens this room and the players whose names are written on that door must enter that room and receive their reward or punishment. (As they do not avoid that room, they do not have to pay their avoidance amounts.)

However, there is a catch. When players declare infinitely high avoidance amounts, it could be the case that all rooms receive a negative infinite net welfare, and this makes room selection impossible. To avoid this problem, the game master sets an upper bound on the sum of the avoidance amounts, which is calculated as follows. Some players are allowed in all the rooms, i.e. their names are on all the doors. These are the ‘necessary players’, because they necessarily receive a payoff. The other players are the ‘contingent players’, because they only receive a payoff based on the contingent fact that their name is written on a door. For each room, the game master calculates the sum of the positive payoffs (rewards) of all the contingent players of that room. The upper bound on the sum of avoidance amounts is given by the maximum of the sums of positive payoffs of contingent players,

[2] A player’s payoff consists of the received reward or punishment, but can also include other considerations that are valued by the player. For example, if a player values equality of rewards, the player may prefer a room with a lower personal reward, if all players in that room receive the same reward. This can be compared with a situation in ethics. Suppose you can choose between two situations. In the first situation, you are very happy (a high utility level), but everyone else is miserable. In the second situation, everyone else becomes extremely happy, at the cost of a slightly lower happiness for you. With some altruistic inclination, you might prefer the second situation, even if you get a lower personal utility.

where the maximum is taken over all the rooms. If there are no rooms with contingent players having positive payoffs, the maximum is simply set to zero.

The analogy with population ethics is as follows. The building corresponds with the choice set in the population ethical problem. Each room corresponds with a possible situation that can be chosen. The rooms where player i is not allowed correspond with the situations where person i does not exist. The players who are allowed in all the rooms, correspond with the necessary people: they exist in all possible situations. The other players correspond with contingent people: their existence depends on the choice of situation. The payoffs are the utilities. A reward U in room s corresponds with a positive utility, i.e. a life worth living in situation s. A punishment corresponds with a negative utility. The declared avoidance amount of a player corresponds with the critical level chosen by a person.

This population ethics game is a strategic interaction between players, because the payoffs of the players depend on the strategies played by the other players. Each player can choose a strategy which consists of his or her declared avoidance amounts for the permissible rooms. As an example, consider a building with three rooms, two necessary players (An and Ben) and thousand contingent players. These contingent players are only allowed in room 3, where they each get a positive but minimal payoff of 1. In room 1, An receives a reward of 300, Ben receives 100. In room 2, the payoffs are reversed: An receives 100, Ben receives 300. In room 3, An and Ben receive a punishment of -100 each. The upper bound on the sum of avoidance amounts is given by the maximum of the sum of positive payoffs of the contingent players, which is 1000. In room 1, Ben can choose a positive avoidance amount of 1000 to maximally influence the selection of room 2, which is his favorite. Similarly, An can choose an avoidance amount of 1000 in room 2. In room 3, An and Ben are worst-off, so they set an infinitely high avoidance amount in order to avoid that situation. However, when they do that, the upper bound of 1000 will be used to calculate the net welfare. Hence, the net welfare values of the rooms 1 and 2 are 400 (the total payoffs of An and Ben) minus 1000 (the total avoidance amount), and for room 3 it is 800 (the total payoffs of An, Ben and the contingent players) minus 1000 (the maximum avoidance amount). As room 3 has the highest net welfare, this room will be chosen. But this room is strongly disliked by An and Ben. We end up with the same conclusion as Sadistic Problem 2.

This conclusion can be avoided: if An and Ben chose a zero avoidance amount, they could manage the selection of their more preferred rooms 1 or 2. By choosing between two strategies, i.e. ‘zero avoidance amount’ and ‘maximum avoidance amount’, An and Ben are in fact playing a strategic game called ‘chicken’ or the ‘hawk-dove game’. This game contains two pure and one mixed Nash equilibria. Both An and Ben playing the strategy ‘maximum avoidance amount’, which results in the conclusion of Sadistic Problem 2, is not a Nash equilibrium.

The rules of variable critical level utilitarianism

The above strategic game resembles population ethics, such that the rules of the game can be translated into the rules of variable critical level utilitarianism. There are five crucial rules.

First, only if a person exists in a situation, that person can choose a non-zero critical level (players can only state a non-zero avoidance amount for the rooms that are allowed for them).

Second, the critical levels cannot be negative.

Third, for each situation the net social welfare is calculated as the difference between total utility and total critical level, or the sum of relative utilities:

$$W(s,\mathcal{C}) = U(s,\mathcal{C}) - C(s,\mathcal{C}) = \sum_{i\in s}\big(u(i,s) - c(i,s,\mathcal{C})\big),$$

where the sum runs over all individuals who exist in situation s.

Fourth, there is a upper bound on the total critical level, given by

$$\bar{C}(\mathcal{C}) = \max_{s\in\mathcal{C}} \sum_{i^c\in s} \max\{u(i^c,s,\mathcal{C}),0\},$$

where the sum runs over all contingent people defined as $i^c \in \{i \,|\, i \in P(s)\ and\ \exists s' \in \mathcal{C}\ with\ i \notin P(s')\}$, i.e. the individual exists in s and there is another situation in which the individual does not exist.

Fifth the optimal situation s_{opt} is chosen as

$$s_{opt}(\mathcal{C}) = \arg\max_{s\in\mathcal{C}}\{W(s,\mathcal{C}), U(s,\mathcal{C}) - \bar{C}(\mathcal{C})\}.$$

There are other ways to construct the upper bound on the total critical level, but the above one is sufficient to avoid the worst counter-intuitive conclusions such as in Sadistic Problem 2. The choice for this upper bound is in a sense natural. First, the upper bound only includes the positive utilities, because adding negative utilities will only decrease the upper bound, moving the theory too much towards total utilitarianism.

Second, the upper bound only includes the utilities of contingent people, because adding the utilities of necessary people would generate the Dependency Problem, where the value of adding extra people depends on how many other people already exist elsewhere or in the past. Suppose all the happy people who ever lived on earth, as well as all the possibly very happy extraterrestrials on far away planets, have to be included in the population of necessary people. After all, their existence cannot be influenced by our choices here and now. Then the upper bound would become very high, and we have to know whether those extraterrestrials really exist. Excluding those previous generations and extraterrestrials decreases the upper bound total critical level, which could change the net social welfare and hence influence the outcome of choosing the most preferred situation.

How variable critical level utilitarianism evades the counterexamples

Let us see how VCLU can avoid the many problems encountered in population ethics. First, it is easy to show that the theory avoids the sadistic conclusions. Consider sadistic conclusion 2, faced by total utilitarianism: in situation s, N people are extremely happy, in situation s' they are extremely miserable, and there are M extra people with lives barely worth living. The N miserable people in situation s' can set very high critical levels, in order to avoid situation s'. If the group of M extra people becomes larger and their total utility increases, the N people can always simply increase their critical levels. Perhaps the extra people who prefer to exist in situation s' can try to counter this, by decreasing their own critical levels, but they cannot decrease them below zero. So the high critical level of the minority group of N people trumps the utilities of the M extra people, which means situation s is chosen.

Also the dominance addition problems can be avoided in some cases. Consider the strong dominance addition problem 2: situation s contains N happy people, M extremely happy people plus a huge number L of extra people with the same happiness as the M-people. Situation s' contains the same N happy people, the same M people who are happier than in situation s, plus a small number Q of

barely happy extra people, not existing in situation s. The M people in situation s can set a high critical level, in order to steer the outcome towards situation s'. However, choosing situation s' may seem counterintuitive to some people, for example in the N population. So, in situation s', those N people can set a very high critical level, to steer the outcome towards situation s. If the M people accept that choosing s' is a bit counterintuitive, they will not raise their own critical levels in situation s. In this case situation s' is avoided and s is selected.

How does the theory deal with the weak dominance addition problem? Situation s contains N very happy people. Situation s' contains the same N people, slightly happier, plus a huge number of M extra, extremely happy people. Situation s'' contains the same N people, as happy as in situation s', plus the same M people as in s' but with lives barely worth living, plus one extra barely happy person. When both situations s' and s'' are possible, the M people in situation s'' can set a high critical level to steer the outcome towards situation s'. The one extra person in situation s'' is not capable to counter this (in order to guarantee existence in situation s''), because that person cannot set a critical level below zero. Also, in situation s, the N people can set a high critical level. Hence, s' can be chosen.

The intransitivity problem of comparativist person-affecting utilitarianism is avoided: the people with the lower utilities can choose the maximum critical levels, and those critical levels in the three situations cancel each other, so the three situations become equally good. When one of the three situations becomes impossible, there arises an inequality between the remaining two situations: the situation where the non-contingent population has the highest utility is the best (where the non-contingent population are the people that exist in both situations).

Conditions that can be violated by variable critical level utilitarianism

VCLU cannot escape the impossibility theorems in population ethics. Here are some conditions that can be violated by VCLU.

> *Quantity Condition* (Arrhenius 2000, p155): For M sufficiently large, situation s with N happy people (and no-one else) is worse than situation s' with no-one except M people who are slightly less happy than the N-people in s.

According to VCLU, it is possible that s is always better than s', no matter how much larger the population M is.

> *Same-Number Condition* (Blackorby, Bossert & Donaldson, 2003): when situations s and s' contain the same number of people and their total utilities are the same, they are equally good.

Yet, the choice set can contain other situations, and it can happen that if the welfare function equalizes s and s', it selects another situation s'' that is really unwanted (sadistic or repugnant). To avoid really unwanted conclusions, if such situations become possible in the choice set, it is possible that the people in s and s' adapt their critical levels such that the social welfare function selects s above s' in this particular choice set. For example, it could be that people in situation s' choose higher critical levels if they wish to avoid the selection of sadistic situations in the choice set that they consider to be far worse. This choice set dependency means a violation of the

> *Choice Set Independency Condition*: whether adding extra people is good or bad should not depend on the choice set, i.e. the possibility to choose other situations.

In general, if people are confronted with a choice set that could result in sadistic or repugnant conclusions, they could choose other critical levels in specific situations, that lead to the violation of some intuitive principles, in order to prevent the selection of a situation that is considered to be worse. VCLU allows to steer away from the most serious counterintuitive implications for each given choice set. This means the theory can also violate e.g. the

> *Situation Independency Condition*: whether adding extra people is good or bad should not depend on e.g. the utilities or the numbers of the already existing people,

as well as the

> *Dominance Addition Condition* (Arrhenius 2000, p159): adding an extra life with positive utility and increasing the happiness of the rest of the population, cannot make a situation worse.

VCLU can also violate the

> *Non-extreme Priority Condition* (Arrhenius, 2000, p163): a situation s that includes N very happy lives and one life with a very small negative utility (close to zero) is always better than a situation s' with N+1 lives with a very low positive utility, other things equal.

This condition can be violated by VCLU when the N people in situation s are not the same as the N people in situation s' (i.e. the N people in s' do not exist in s). But if the one person in situation s has only a very small negative utility and a population ethical preference for situation s', that person can choose a zero critical level, such that situation s can still be selected.

In summary, if worrisome conclusions do not have to be avoided, VCLU can easily respect the abovementioned conditions.

Dynamic inconsistency

Another implication of VCLU, is the possibility of dynamic inconsistency (when a game is subgame imperfect; see Simaan and Cruz (1973) for original work, and Kydland & Prescott (1977) for other famous examples of dynamic inconsistencies in e.g. macroeconomics). Consider a choice set with three situations. Situation s contains N very happy people, with a high average per capita utility level $U_N(s)$. Situation s' contains the same N people, who are slightly happier (utility $U_N(s')>U_N(s)$) plus M extra people with low happiness (positive utility $U_M(s')<U_N(s)$). Situation s'' contains the same N people, slightly less happy than in situation s (utility $U_N(s'')<U_N(s)$), and the extra M people who are very happy (utility $U_M(s'')>U_M(s')$).

The game consists of two choices or stages. The first choice involves the addition of the extra M people. The second stage occurs once the M people are chosen to be added, and involves choosing low or high utilities for those M people (i.e. situations s' or s''). This game can be solved with backward induction, where we first consider the final subgame, i.e. the stage when the M people are chosen to be added. Although situation s' is the best for the N people, the M people in that situation can complain and prefer situation s'', such that they choose maximum critical levels totaling $M.U_M(s'')$. In situation s'', the N people can complain and set maximum critical level also totaling $M.U_M(s'')$, to turn the balance again in favor of situation s'. The critical levels cancel, so the situation with the highest total utility will be chosen. Suppose $N.U_N(s'')+M.U_M(s'')>N.U_N(s')+M.U_M(s')$, then situation s'' is chosen. However, this solution for the subgame is not an equilibrium in the complete game, because situation s is preferred to situation s'' by the N people. In other words: the N people

cannot accept the existence of the M people, because if they did so, they know that the end result will be a situation s'' that has a lower payoff than the situation s without the M people. If the N people choose a maximum critical level in s, then they know the selected situation will be s'', which they do not prefer. Therefore, they can choose to set a low or critical level in s, which means in the first stage of the game situation s will be selected. What is optimal in a subgame where the choice set consists of s' and s'', becomes suboptimal in the complete game with choice set {s,s',s''}. The optimal choice depends on the stage in the game. This is known as dynamic inconsistency.

Here we see again that VCLU is choice set dependent. If situation s'' was not possible (i.e. was not an element of the choice set), situation s' could become better than s (because the M people in s' can no longer complain that situation s'' should have been chosen). The value of adding extra people depends on the possible situations that contain those people.

Examples of dynamic inconsistency

The dynamic inconsistency of VCLU can be explained with some examples. First, consider a more than hundred year old argument in favor of meat consumption, called the 'Logic of the larder' (Salt, 1914; see also Matheny & Chan, 2005). This argument concerns 'happy meat', i.e. meat from a livestock animal that had a life worth living (a net-positive life with more positive than negative experiences). In situation s, meat consumption and livestock farming are not allowed and N humans need to eat vegan food. In situation s', animals are raised at happy farms (no factory farms) where they have net-positive welfare, but they are killed prematurely so that humans become a little happier by enjoying the taste of meat. In situation s'', those animals are not killed prematurely, but can live long happy lives at farm animal sanctuaries. Their happiness increases a lot, but now humans can no longer eat meat, and they have to take care of the animals (e.g. feeding them), which bears an extra cost. In this situation, humans get the lowest welfare (lower than in situation S), but still positive.

Salt (1914) argued that eating happy meat (situation s') is not allowed, by comparing the situation with human slavery: we are not allowed to breed human slaves, even if those slaves would have net-positive lives. It is better that those happy human slaves are not born, so Salt prefers situation S. This is also the outcome of VCLU, due to the dynamic inconsistency.

Suppose the happy livestock animals or happy human slaves had such positive lives, that they prefer existence (as meat animals or slaves) above non-existence. When situation s'' is part of the choice set, those animals or slaves could complain once they exist in situation s'. However, if they would complain, the already existing N humans would decide not to breed those people, because they want to avoid situation s''. However, if it would be possible to exclude situation s'' from the choice set, situation s' could be chosen (by choosing lower critical values in s'). In games with dynamic inconsistency, this can be done with a commitment device (Brocas e.a. 2004). Suppose for example that we can genetically modify a cow such that the cow will die at the age of two years (when he normally gets slaughtered in situation s'). The cow can be raised on a farm sanctuary, and after the cow dies, he can be eaten.

Another example of dynamic inconsistency is climate change. In situation s, the current generation (N people) invest enough in climate policies and clean energy such that harmful climate change is avoided and the next generation (L people) have very happy lives. In situation s', the current generation does nothing about climate change, they are happier because they can consume more and worry less, but their decision to travel a lot with cars and airplanes influences the exact timing of

fertilization of their future children. Having sex a second later, and a son instead of a daughter is born. As a consequence, the next generation is not the L people, but other people are born. These M people do not exist in situation s. Suppose the M people in situation s' have to deal with the consequences of dangerous climate change, but they still have slightly positive lives. These people prefer a third situation s'', where they exist and get huge compensation fees from the N people who caused climate change. In s'', the M people are happier, but due to the compensation payments, the N people become worse off in s'' than in situation s. In this case, variable critical level utilitarianism could pick situation s.

This consideration also influences the discount rate that is used in cost-benefit analyses of climate policies. If situation s is chosen, it implies that the welfare of the next generation should not be discounted much: it is better that the next generation is very happy (the L people in situation S) instead of slightly happy (the M people in situation s'). However, the welfare of generations in the more distant future can be strongly discounted according to variable critical level utilitarianism. For the more distant future, this population ethical theory can resemble the asymmetric person-affecting theories. This is because in the more distant future, the current generation no longer exists and hence is no longer able to pay compensation fees to e.g. the Q people of the fifth generation. Suppose those Q people had low but still positive welfare levels due to climate change. They cannot complain against the N people (the current generation), because if the N people chose policies to avoid climate change, the Q people would not be born. In other words, a situation analogous to s'' for the Q people in the more distant future is impossible. That means the welfare of further generations in the more distant future can be strongly discounted (at least when they still have positive utilities: when they get a negative utility due to climate change, their negative relative utilities strongly decrease the welfare function because their critical levels cannot go below zero).

Conclusion

Variable critical level utilitarianism is a theory in population ethics that uses a welfare function composed of the sum of the relative utilities of all existing people, whereby a relative utility is the actual utility of a person in a situation, minus a critical level. These critical levels are variable: people are free to choose their own critical levels (up to a well-chosen maximum), and so these critical levels can differ between situations and can even depend on the choice sets of possible situations. Traditional population ethical theories are limiting cases of variable critical level utilitarianism, with constraints on the critical levels. Due to these restrictions of the critical levels, those traditional theories face counterintuitive implications such as sadistic and repugnant conclusions.

The flexibility of variable critical level utilitarianism allows to avoid the population ethical problems. The fact that people can choose their own critical level and take into account the choices of other existing people, creates a strategic game. Variable critical level utilitarianism has a game theoretic dynamic inconsistency. Some examples (consuming meat from happy livestock animals, breeding happy human slaves, causing climate change) demonstrate that this dynamic inconsistency is a virtue rather than an vice: it can explain when and why breeding happy livestock animals or happy human slaves is not allowed, why we have to prevent climate change and why we should not strongly discount the welfare of at least the next few generations.

In future research, the game theoretic implications of variable critical level utilitarianism can be investigated further. For example, what are the effects of coordination and cooperation, when people can collectively choose their critical levels? Also, given the choice set of possible future

trajectories of our society, what critical levels do people choose, and how does this influence the discount rate and the balancing of current versus future welfare?

References

Arrhenius, G. (2000). *Future Generations: A Challenge for Moral Theory*, PhD dissertation, Uppsala University.

Arrhenius, G. (2000b). An impossibility theorem for welfarist axiologies. *Economics and Philosophy*, *16*(2):247-266.

Arrow, K. (1950). A Difficulty in the Concept of Social Welfare. *Journal of Political Economy*. 58 (4): 328–346.

Blackorby, C., Bossert, W., and Donaldson, D. (1995). Intertemporal Population Ethics: Critical Level Utilitarian Principles, *Econometrica* 65:1303-1320.

Blackorby, C., Bossert, W., and Donaldson, D. (1997). Critical-Level Utilitarianism and the Population-Ethics Dilemma. *Economics and Philosophy*, 13:197-230.

Blackorby, C., Bossert W., & Donaldson, D. (2002). Population Principles with Number-Dependent Critical Levels, *Journal of Public Economic Theory,* 4:347–68.

Blackorby, C., Bossert W., & Donaldson, D. (2003). The Axiomatic Approach to Population Ethics. *Politics Philosophy Economics*, 2(3): 342-381.

Blackorby, C., Bossert W., & Donaldson, D. (2005). *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*. Cambridge University Press.

Brocas, I., Carrillo, J. D., & Dewatripont, M. (2004). Commitment devices under self-control problems: An overview. *The Psychology of economic decisions*, *2*, 49-67.

Bykvist, K. (1998). *Changing Preferences: A Study in Preferentialism*, F.D. Diss., Uppsala University.

Dasgupta, P. (1988). Lives and Well-Being, *Social Choice and Welfare* 5:103-126.

Fehige, C. (1992). A Pareto Principle for Possible People. Univ. of Pittsburgh, Center for Philosophy of Science.

Heyd, D. (1988). Procreation and Value: Can Ethics Deal With Futurity Problems? *Philosophia*18:151-170.

Hurka, T. (1983). Value and Population Size. *Ethics,* 93: 496–507.

Kydland, F. & Prescott, E. (1977). Rules Rather than Discretion: The Inconsistency of Optimal Plans. *Journal of Political Economy*. 85 (3): 473–492.

Matheny, Gaverick & Kai Chan. 2005. Human diets and animal welfare: the illogic of the larder. *Journal of Agricultural and Environmental Ethics* 18(6): 579-594.

Meacham, C. J. (2012). Person-affecting views and saturating counterpart relations. *Philosophical Studies*, *158*(2), 257-287.

Mulgan, T. (2006). *Future People. A Moderate Consequentialist Account of our Obligations to Future Generations*, Oxford: Clarendon Press.

Myerson, R. & Satterthwaite, M. (1983). Efficient Mechanisms for Bilateral Trading. *Journal of Economic Theory*. 29 (2): 265–281.

Narveson, J. (1973). Moral Problems of Population. *The Monist* 57:62-86.

Ng, Y-K. (1989). What Should We Do About Future Generations? Impossibility of Parfit's Theory X. *Economics and Philosophy* 5(2): 235-253.

Parfit, D. (1984). *Reasons and Persons*, Oxford: Clarendon Press.

Salt, H. (1914). Logic of the Larder. *The Humanities of Diet, Manchester*, 221-222.

Simaan, M. & Cruz, J. (1973). On the Stackelberg Strategy in Nonzero-Sum Games. *Journal of Optimization Theory and Applications*. 11 (5): 533–555.
Shields, L. (2012). The prospects for sufficientarianism. *Utilitas*, *24*(1), 101-117.
Walker, A. D. M. (1974). Negative utilitarianism. *Mind*, *83*(331), 424-428.