Protected Income and Inequality Aversion*

Marc Fleurbaey† and Eduardo Zambrano‡

February 18, 2025

Abstract

We discover a fundamental and previously unrecognized structure within the class of additively separable social welfare functions that makes it straightforward to fully characterize and elicit the social preferences of an inequality-averse evaluator. From this structure emerges a revealing question: if a large increment can be given to one individual in a society, what is the maximal sacrifice that another individual can be asked to bear for its sake? We show that the answer uncovers the evaluator's degree of inequality aversion. In particular, all translation-invariant evaluators would sacrifice the full income of the sacrificed individual if their income were low enough and a declining fraction of their income otherwise. Scale-invariant evaluators would sacrifice the full income of the sacrificed individual at all income levels if their inequality aversion was no greater than one, and a constant fraction of their income otherwise. Motivated by these findings, we propose a class of social preferences that, starting from a baseline level of protection, ensure a higher fraction of the sacrificed individual's income is protected the lower their income.

*The authors warmly thank audiences at workshops and seminars at CERE-Umeå, ISER-Osaka, MDOD-Sorbonne, and Cal Poly San Luis Obispo for very helpful discussions. The authors also expresses gratitude to the Princeton School of Public and International Affairs for hosting Zambrano's sabbatical and to the Paris School of Economics for hosting Zambrano during the final stages of this research. This research has been supported by the National Science Foundation (USA) under Grant # 2313969.

†Paris School of Economics, Centre National de la Recherche Scientifique, and École normale supérieure - PSL, France.

‡Corresponding Author. Department of Economics. Cal Poly - San Luis Obispo. USA.

Keywords: Social Welfare Functions; Inequality Aversion; Revealed Ethics.

JEL Classification: D31, D63, D90.

1 Introduction

Since Atkinson's (1970) and Kolm's (1976) seminal works on inequality measurement, the notion of inequality aversion has been well understood, and applied to many different contexts. The concept of inequality aversion has been closely associated to the Pigou-Dalton transfer principle, according to which social welfare does not decrease (or inequality does not increase) when a given amount is taken from someone and given to someone else who is less well-off, without reversing the ranking between these two individuals.[1]

The empirical-experimental literature on people's social preferences[2] has elicited the value of their inequality aversion about income by making respondents choose between distributions involving different levels of inequality and different levels of average income. For instance, Venmans and Groom (2021) find an average coefficient of inequality aversion of 2.9 in an empirical survey on a sample of students, while Hurley et al. (2020) find an average coefficient of 3.27 in a representative sample of the Ontario province of Canada. The underlying mechanism for such ethical preference elicitation is the leaky-bucket experiment: a distribution is less unequal than another, but has a lower average income, reflecting the classical equality-efficiency trade-off highlighted in Okun (1975). The meaning of the coefficient of inequality aversion is usually presented in this way: a coefficient of 3 means that it is worth giving a dollar to someone even if this requires taking 8 ($= 2^3$) dollars from a person who is twice as rich—thus losing 7 dollars, i.e., 87.5% of the amount levied from the donor.

A small literature (Fleurbaey and Michel 2001, Aboudi and Thon 2003, Thon and Wallace 2004, Dubois 2016) has explored the relationship between principles of transfer that allow for leakage and the value of the coefficient of inequality aversion in a social welfare criterion that endorses such leaky trans-

[1] Aboudi and Thon (2006) also study what they call the Muirhead-Dalton principle, in which transfers are allowed to reverse the ranking between the two individuals, but still leave the difference between them smaller than in the initial distribution. For authoritative reviews of this field, see Chakravarty (2009), Dutta (2002) and Cowell (2000).

[2] For a review, see Gaertner and Schokkaert (2012).

fers.[3] For instance, Fleurbaey and Michel (2001) show that if the transfers are proportional to the initial level of the donor and recipient, the coefficient must be at least 2, whereas if the transfers are proportional to the final levels, the coefficient must be at least 1.

A coefficient of 2, according to the common presentation (referring to a gift of 1 dollar costing $2^2 = 4$) means that wasting 75% of the transfer when the recipient has only half the level of the donor is acceptable, and this seems a lot. The principle of proportional transfers (on initial levels), however, suggests something less extreme, since when the recipient gets X, a donor who is *initially* twice as rich would only lose 2X, which means that only half of the gift is lost in the transfer. This apparent contradiction is resolved by noting that the reasoning in terms of one dollar gifts is only true for infinitesimal gifts, whereas the principe of proportional transfers applies to transfers of any size (under the proviso that they do not reverse the order between the donor and the recipient).

This shows that the way in which the implications of the coefficient of inequality aversion are presented may influence the ethical choice of the coefficient. One may find that the principle of proportional transfers appears reasonable, since it is very common for people to think in terms of variations in percentage, while hesitating to condone the maxim that "a marginal transfer from a donor to someone who is half below can waste 75% and still improve social welfare"—and yet the two refer to the same coefficient of inequality aversion of 2.

In this paper we propose to study an intuitive and tractable framing of the problem of choosing the coefficient of inequality aversion. It refers to the following narrative. Start from a perfectly equal distribution, and then imagine that an increment can be given to one individual. What is the maximal sacrifice that can be imposed on another individual for the sake of this increment? In other words, if one individual's level of income can become infinitely high, how much could justifiably be taken from another one in order to make this possible? We propose a formal analysis of this story in a setting with a separably additive social welfare function. It seems to us that this framing pushes people's intuition further toward high values of inequality aversion. In particular, a coefficient that is below 1 would justify depriving the sacrificed individual from everything. For a Kolm-Atkinson social welfare

[3]See also Bosmans (2014) for a similar analysis applied to rank-dependent poverty measures, involving transfers across the poverty threshold.

function, a coefficient of 2 or more means that it is never acceptable to bring the sacrificed individual below half of her initial level, which seems a rather generous (i.e., generously *low*) bound—the bound is 70% of the initial level when the coefficient is 3. In addition to relating levels of protection of the sacrificed individual's income to coefficients of inequality, we actually characterize the classes of additively separable social welfare functions that guarantee specific (absolute or relative) levels of protection.

The paper is structured as follows. In section 2, we define the framework and the main notions. Section 3 determines the general conditions under which an additively separable social welfare function provides a strictly positive protected level to the sacrificed individual. In section 4, we characterize the class of social welfare functions which protect a fixed fraction of the initial level, against increments given to either one other individual, or two other individuals. Section 5 instead examines the case in which the protected level is always at a fixed absolute difference from the initial level. Section 6 explores how to obtain a protection that is, in proportion to the initial level, greater when the initial level is lower, reflecting the intuitively appealing thought that poor people deserve a relatively greater protection than rich people. We examine in section 7 the case in which the starting point is not perfect equality and generalize the results to this broader class of contexts. Section 8 concludes. We also created a companion website to the paper, available at https://osf.io/tnu2q/, which contains interactive visuals, apps and narratives that explain and motivate our results in simple and intuitive terms.

2 Setting and main definitions

Consider a population of n individuals, each individual i enjoying income $y_i \geq 0$, and an additively separable social welfare function

$$W(y_1, ..., y_n) = \sum_{i=1}^{n} f(y_i),$$

where f is an increasing function with values taking an interval in $\mathbb{R} \cup \{-\infty, +\infty\}$. Throughout this paper we restrict attention to this class of social welfare functions.

Income is used as the index of individual advantage in this paper, but obviously the analysis is relevant to any setting in which individual advan-

tage is measured by a cardinally measurable and interpersonally comparable index. In particular, it is possible to adjust income for non-market aspects of quality of life that individuals enjoy or endure, and use this ajusted income (usually called "equivalent income") as the relevant index instead of ordinary income.

The social welfare function is symmetric, which reflects impartiality among individuals. Therefore, we can focus on individual 1 (called Ana, to fix ideas) in trade-offs with any number of fellow individuals.

Consider a starting point in which Ana and individual 2 (called Ben) have the same income y. Then one can consider unequal allocations that keep social welfare constant, i.e., spans allocations (y_1, y_2) such that

$$f(y_1) + f(y_2) = 2f(y).$$

In particular, focus on how Ana's income decreases when Ben's in raised. One can define the function $y_1(y_2, y)$ implicitly by the equation

$$f(y_1(y_2, y)) + f(y_2) = 2f(y),$$

yielding the explicit form

$$y_1(y_2, y) = f^{-1}(2f(y) - f(y_2)).$$

The function $y_1(y_2, y)$ is increasing in y and decreasing in y_2, because f is increasing and so is f^{-1}.

The domain of definition of $y_1(y_2, y)$ depends on f. For instance, if $f(y) = y$, then $y_1(y_2, y)$ is defined only for (y_2, y) such that $y_2 \leq 2y$. Let D_{12} denote the domain of definition of $y_1(y_2, y)$:

$$D_{12} = \{(y_2, y) \,|2f(y) - f(y_2) \geq f(0)\}.$$

We can now introduce the notion of *protected income*, which is the level of income below which y_1 will never fall when y_2 increases. No matter what opportunities for high income are offered to Ben, Ana's income will never fall below this level if social welfare is preserved.

Protected income: For a given $y > 0$, the level of protected income is defined as

$$\ddot{Y}(y) = \inf\{y_1(y_2, \hat{y}) \,|\, (y_2, \hat{y}) \in D_{12}, \hat{y} = y\}.$$

Relatedly, we define *collateral damage* as the sacrifice that can be imposed on Ana for the sake of raising Ben's income.

Collateral damage: For a given $y > 0$, the level of collateral damage is defined as

$$\ddot{L}(y) = y - \ddot{Y}(y).$$

The *relative* collateral damage is

$$\frac{\ddot{L}(y)}{y} = 1 - \frac{\ddot{Y}(y)}{y}.$$

3 Existence of a positive protected income

Not all social welfare functions guarantee Ana against complete sacrifice for the sake of Ben's stratospheric income. Actually, many of the most commonly used social welfare functions do not protect against complete sacrifice.

Proposition 1. *For all $y > 0$, one has $\ddot{Y}(y) > 0$ if and only if f has an upper bound strictly below $2f(y) - f(0)$.*

Proof. Let y be given. If f has an upper bound strictly below $2f(y) - f(0)$, then there is ε such that for all $y_2 \geq 0$,

$$f(y_2) < 2f(y) - f(0) - \varepsilon,$$

implying that if $f(y_1) + f(y_2) = 2f(y) > f(0) + f(y_2) + \varepsilon$, then necessarily $f(y_1) > f(0) + \varepsilon$, so that y_1 is bounded away from zero. In this case, $\ddot{Y}(y) > 0$.

If f does not have an upper bound strictly below $2f(y) - f(0)$, then for all ε, there is y_2 such that

$$f(y_2) > 2f(y) - f(0) - \varepsilon,$$

implying that if $f(y_1) + f(y_2) = 2f(y) < f(0) + f(y_2) + \varepsilon$, then necessarily $f(y_1) < f(0) + \varepsilon$, so that y_1 is not bounded away from zero. In this case, $\ddot{Y}(y) = 0$. □

To illustrate this result, consider the class of Kolm-Atkinson social welfare functions, for which

$$f(y) = \begin{cases} \frac{y^{1-\eta}}{1-\eta} & \text{for} \quad \eta \neq 1, \eta \geq 0 \\ \ln y & \text{for} \quad \eta = 1 \end{cases}$$

where $\eta \geq 0$ is the coefficient of inequality aversion.

When $\eta \leq 1$, f has no upper bound and thus $\ddot{Y}(y) = 0$ for all $y > 0$. When $\eta > 1$, one has $f(0) = -\infty$, so that $2f(y) - f(0) = +\infty$, whereas f is always negative. Thus, $\ddot{Y}(y) > 0$ for all $y > 0$. More precisely, one has

$$\ddot{Y}(y) = f^{-1}(2f(y)) = 2^{\frac{1}{1-\eta}} y.$$

This is a quite significant result because values of $\eta = 0.5$ and 1 are frequent in the applied literature on inequalities. With such values, the ultimate sacrifice of giving all of one's income for the sake of a billionaire is considered acceptable, and this appears hard to defend. If one would like to avoid such a repugnant conclusion, it is imperative to pick $\eta > 1$.

Another important class is formed by the Kolm-Pollak social welfare functions, for which

$$f(y) = \begin{cases} -e^{-\alpha y} & \text{for} \quad \alpha > 0 \\ y & \text{for} \quad \alpha = 0 \end{cases}$$

where $\alpha \geq 0$ is the coefficient of inequality aversion. When $\alpha = 0$, one obviously has $\ddot{Y}(y) = 0$ for all $y > 0$. When $\alpha > 0$, f is negative with upper bound zero, whereas $2f(y) - f(0)$ is non-positive when $y \leq \frac{\ln 2}{\alpha}$, and thus $\ddot{Y}(y) = 0$ for all $y \leq \frac{\ln 2}{\alpha}$. When $y > \frac{\ln 2}{\alpha}$, one has

$$\ddot{Y}(y) = f^{-1}(2f(y)) = y - \frac{\ln 2}{\alpha}.$$

Unlike the Kolm-Atkinson class, members of the Kolm-Pollak class never guarantee a positive protected level of income for all values of y, no matter how high the coefficient of inequality aversion is. This can be seen as a serious disadvantage for this class, although the level of y below which the protected income falls to zero becomes vanishingly small when α is high.

A graphical representation of $\ddot{Y}(y)$ is easy to build upon the graph of f. Given that f must have an upper bound, there is no loss of generality to let $f(+\infty) = 0$. One then has $\ddot{Y}(y) = f^{-1}(2f(y))$. This is illustrated in Fig. 1.

4 Protecting a fraction of income

Consider again the Kolm-Atkinson social welfare functions, for the relevant case $\eta > 1$. Since

$$\ddot{Y}(y) = 2^{\frac{1}{1-\eta}} y,$$

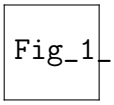


Figure 1: Representation of $\ddot{Y}(y)$

the protected income is always a fixed fraction of the initial (equal) income. The fraction $2^{\frac{1}{1-\eta}}$ tends to zero when η is close to 1, and tends to one when η is very high. So, this fraction spans the whole range, which connects the coefficient of inequality aversion to the protected income in a convenient way.

The next proposition characterizes the class of social welfare functions protecting a fixed fraction of income.

Proposition 2. *The subclass of social welfare functions satisfying, for all $y > 0$ and a fixed $\lambda \in (0,1)$, the property $\ddot{Y}(y) = \lambda y$ is generated by the function*

$$f(y) = -e^{g(\ln y)},$$

where g is defined as follows. It is an arbitrary continuous decreasing function over the interval $[0, -\ln\lambda]$ such that $g(\ln\lambda) - g(0) = \ln 2$, and over the rest of $\mathbb{R}$ it is defined in a periodic way, i.e., $g((k+1)\ln\lambda + a) - g(k\ln\lambda + a) = \ln 2$ for all $k \in \mathbb{Z}_+$,$a \in [0, -\ln\lambda]$.

Proof. From Prop. 1, f must have an upper bound. There is no loss of generality in positing $f(+\infty) = 0$. (Note that f is thus negative.) One then has $\ddot{Y}(y) = f^{-1}(2f(y))$.

We are looking for the class of f satisfying, for all $y > 0$:

$$f^{-1}(2f(y)) = \lambda y$$

for a fixed λ. This can be written as

$$2f(y) = f(\lambda y).$$

Letting $g = \ln(-f \circ e)$, a change of variables $x = \ln y$ (note that g is decreasing and is defined on $\mathbb{R}$) yields

$$\ln 2 + g(x) = g(\ln\lambda + x).$$

This property is satisfied by any function g such that for all $k \in \mathbb{Z}_+$,

$$g(k\ln\lambda) = g(0) + k\ln 2$$

(this defines a line of points of intercept $g(0)$ and negative slope $\ln 2/\ln\lambda$) and g is, suitably detrended, periodic across intervals $[(k+1)\ln\lambda, k\ln\lambda]$, i.e., for $a \in [0, -\ln\lambda]$,

$$g(k\ln\lambda + a) + \ln 2 = g((k+1)\ln\lambda + a).$$

Once such a function has been picked, one then retrieves

$$f(y) = -e^{g(\ln y)}.$$

□

Since the graph of g contains a line of points, the case of a linear g is obviously salient. One then has $g(x) = x\ln 2/\ln\lambda$, and

$$f(y) = -y^{\ln 2/\ln\lambda},$$

which brings us back to the Kolm-Atkinson class, for $\eta = 1 - \ln 2/\ln\lambda$, and this implies that the fraction of protected income is indeed $2^{\frac{1}{1-\eta}}$.

Let us now extend the notion of protected income to the case in which individuals 2 and 3 see their incomes rise without limit. Let the function $y_1(y_2, y_3, y)$ be defined implicitly by the equation

$$f(y_1(y_2, y)) + f(y_2) + f(y_3) = 3f(y),$$

and explicitly as

$$y_1(y_2, y_3, y) = f^{-1}(3f(y) - f(y_2) - f(y_3)).$$

It is increasing in y and decreasing in both y_2, y_3.

The domain of definition of $y_1(y_2, y_3, y)$, denoted D_{123} is defined as follows

$$D_{12} = \{(y_2, y_3, y) \,|3f(y) - f(y_2) - f(y_3) \geq f(0)\}.$$

Protected income against two: For a given $y > 0$, the level of protected income against two other individuals is defined as

$$\dddot{Y}(y) = \inf\{y_1(y_2, y_3, \hat{y}) \,|\, (y_2, y_3, \hat{y}) \in D_{123}, \hat{y} = y\}.$$

In the case of the Kolm-Atkinson class, one obtains

$$\dddot{Y}(y) = 3^{\frac{1}{1-\eta}} y,$$

so that once again, a fixed fraction of y is guaranteed to Ana. If one combines the requirement to protect a fixed (different) fraction of income against one and two other individuals, one obtains a characterization of this class (up to an affine transform).

Proposition 3. *The Kolm-Atkinson subclass of social welfare functions is (up to an affine transform) characterized by the requirement that there exist* $\lambda, \mu \in (0,1)$ *such that for all* $y > 0$, $\ddot{Y}(y) = \lambda y$ *and* $\dddot{Y}(y) = \mu y$. *One must then have* $\lambda = 2^{\frac{1}{1-\eta}}$ *and* $\mu = 3^{\frac{1}{1-\eta}}$ *for some* $\eta > 1$.

Proof. From Prop. 2 we know that $f(y) = -e^{g(\ln y)}$ for a function g such that

$$g((k+1)\ln\lambda + a) - g(k\ln\lambda + a) = \ln 2$$

for all $k \in \mathbb{Z}_+$,$a \in [0, -\ln\lambda]$. By a similar reasoning, the function g must also satisfy

$$g((k+1)\ln\mu + a) - g(k\ln\mu + a) = \ln 3$$

for all $k \in \mathbb{Z}_+$,$a \in [0, -\ln\mu]$. In particular, for all $k \in \mathbb{Z}_+$,

$$g(k\ln\lambda) - g(0) = k\ln 2$$

$$g(k\ln\mu) - g(0) = k\ln 3.$$

Given the quasi-periodic behavior of g, there exist $b < 0 < c$ such that for all $x \in \mathbb{R}$,

$$b < g(x) - g(0) - x\frac{\ln 2}{\ln\lambda} < c$$

$$b < g(x) - g(0) - x\frac{\ln 3}{\ln\mu} < c.$$

This is possible only if

$$\frac{\ln 2}{\ln\lambda} = \frac{\ln 3}{\ln\mu}.$$

Let $1-\eta$ be this common (negative) ratio. This proves that $\lambda = 2^{\frac{1}{1-\eta}}$ and $\mu = 3^{\frac{1}{1-\eta}}$ for some $\eta > 1$.

Without loss of generality (i.e., up to an affine transform of W), we can let $f(+\infty) = 0$ and $f(1) = \frac{1}{1-\eta}$. Then

$$\begin{cases} f\left(2^{\frac{1}{1-\eta}}y\right) = 2f(y) \\ f\left(3^{\frac{1}{1-\eta}}y\right) = 3f(y) \end{cases}$$

The rest of the argument follows the proof of the first half of Theorem A in Tossavainen and Haukkanen (2007). Let $S = \left\{ 2^{\frac{m}{1-\eta}}3^{\frac{n}{1-\eta}} \,\middle|\, m, n \in \mathbb{Z} \right\}$ and pick $y \in S$. Then,

$$f(y) = f\left(2^{\frac{m}{1-\eta}} 3^{\frac{n}{1-\eta}}\right) = 2^m f\left(3^{\frac{n}{1-\eta}}\right) = 2^m 3^n f(1)$$
$$= \left(2^{\frac{m}{1-\eta}} 3^{\frac{n}{1-\eta}}\right)^{1-\eta} f(1) = f(1)\, y^{1-\eta} = \frac{y^{1-\eta}}{1-\eta}.$$

We have established that $f(y) = \frac{y^{1-\eta}}{1-\eta}$ for $y \in S$. In order to extend the result to all y in the interval $(0, +\infty)$ we need to show that S is dense in $(0, +\infty)$. Pick y in $(0, +\infty)$ and let $T = \{m + n \log_2 3 | \, m, n \in \mathbb{Z}\}$. Since $\log_2 3$ is irrational this means that T is dense in $\mathbb{R}$. Since $(1-\eta)\log_2 y \in \mathbb{R}$, there is a sequence $\{t_j\} \subset T$ such that

$$\lim_{j\to\infty} t_j = (1-\eta)\log_2 y,$$

which implies that $\lim_{j\to\infty} \frac{t_j}{1-\eta} = \log_2 y$, or $\lim_{j\to\infty} 2^{\frac{t_j}{1-\eta}} = y$. The next step is to show that $\left\{2^{\frac{t_j}{1-\eta}}\right\} \subset S$. This is readily seen, since $2^{\frac{t_j}{1-\eta}} = 2^{\frac{m+n\log_2 3}{1-\eta}} = 2^{\frac{m}{1-\eta}} \left(2^{\log_2 3}\right)^{\frac{n}{1-\eta}} = 2^{\frac{m}{1-\eta}} 3^{\frac{n}{1-\eta}}$ for some $m, n \in \mathbb{Z}$, which means that $2^{\frac{t_j}{1-\eta}} \in S$ for each j. It follows that S is dense in $(0, +\infty)$ and therefore that $f(y) = \frac{y^{1-\eta}}{1-\eta}$ for all $y \in (0, +\infty)$, which is what we wanted to show. □

5 Constant Collateral Damage

We have criticized the Kolm-Pollak social welfare functions for not guaranteeing a positive income when the initial income is low. But it is worth revisiting the property of a constant collateral damage, i.e., a protected income that is always at the same distance from the initial income, provided the latter is greater than this constant collateral damage.

Here again, we assume that f has an upper bound and that $f(+\infty) = 0$. The next proposition characterizes the class of social welfare functions exhibiting constant collateral damage.

Proposition 4. *The subclass of social welfare functions satisfying, for a fixed $\Delta \in (0, +\infty)$ and all $y > \Delta$, the property $\ddot{Y}(y) = y - \Delta$ is generated by the function*

$$f(y) = -e^{g(y)},$$

where g is defined as follows. It is an arbitrary continuous decreasing function over the interval $[0, \Delta]$ such that $g(\Delta) - g(0) = -\ln 2$, and over the rest of

$(0, +\infty)$ *it is defined in a periodic way, i.e.,* $g((k+1)\Delta + a) - g(k\Delta + a) = -\ln 2$ *for all* $k \in \mathbb{Z}_+$, $a \in [0, \Delta]$.

Proof. Constant collateral damage means that for all $y > \Delta$,

$$f(y - \Delta) = 2f(y).$$

Let $g = \ln f$. For all $y > 0$, one has

$$g(y + \Delta) = g(y) - \ln 2.$$

For $y = 0$, one obtains $g(\Delta) - g(0) = -\ln 2$, and for $y = k\Delta + a$, one obtains $g((k+1)\Delta + a) - g(k\Delta + a) = -\ln 2$. □

In similar fashion as in the previous section, one can add the requirement that constant collateral damage also occurs against two other individuals, and derive a characterization of the Kolm-Pollak class of social welfare functions.

Proposition 5. *The Kolm-Pollak subclass of social welfare functions is (up to an affine transform and above a minimum level of income) characterized by the requirement that there exist* $\Omega > \Delta > 0$ *such that for all* $y > \Omega$, $\ddot{Y}(y) = y - \Delta$ *and* $\dddot{Y}(y) = y - \Omega$. *One must then have* $\Delta = \frac{\ln 2}{\alpha}$ *and* $\Omega = \frac{\ln 3}{\alpha}$ *for some* $\alpha > 0$.

Proof. See Appendix 10.1. □

6 Progressive Protection

The proportional protected income, with $\eta > 1$, is an improvement over the constant collateral damage in the sense that in the latter case, collateral damage, in proportion to income, is greater at lower levels of income. However, the fact that *relative* collateral damage is indeed constant with the Kolm-Atkinson approach, no matter what level of income one is considering, may be viewed as problematic, too. One might want to consider orderings that protect a higher fraction of Ana's income (and therefore allow lower relative collateral damage) the lower her income is. We turn our attention now to a novel class of orderings where this would be so.

Fix a parameter $c > 0$. Let f be defined, for $y \geq c$, by:

$$f(y) = \begin{cases} \frac{\left(\ln \frac{y}{c}\right)^{1-\gamma}}{1-\gamma} & \text{for} \quad \gamma \neq 1, \gamma \geq 0 \\ \ln \ln \frac{y}{c} & \text{for} \quad \gamma = 1. \end{cases}$$

We first note that, once again, $\gamma \leq 1$ does not embed enough inequality aversion in that, for those social welfare functions, there is no substantial protection. This follows directly from the fact that f has no upper bound in this case, so that $2f(y) - f(y_1)$ can be passed by $f(y_2)$ for y_2 great enough, no matter how low $y_1 > c$ is.

On the other hand, for $\gamma > 1$, these social welfare functions exhibit the feature that a higher fraction of income is protected the lower Ana's income is and therefore the orderings have increasing relative collateral damage.

Proposition 6. *Consider the subclass of social welfare functions generated by f as defined above. Then, for all $y > c$,$\ddot{Y}(y) = y^{\ddot{\beta}} c^{1-\ddot{\beta}}$, where $\ddot{\beta} = 2^{\frac{1}{1-\gamma}}$.*

Proof. We compute that

$$y_1(y_2; y) = c \cdot e^{\left[2\left(\ln \frac{y}{c}\right)^{1-\gamma} - \left(\ln \frac{y_2}{c}\right)^{1-\gamma}\right]^{\frac{1}{1-\gamma}}}$$

Notice that $\lim_{y_2 \to +\infty} \left[2\left(\ln \frac{y}{c}\right)^{1-\gamma} - \left(\ln \frac{y_2}{c}\right)^{1-\gamma}\right]^{\frac{1}{1-\gamma}} = 2^{\frac{1}{1-\gamma}} \ln \frac{y}{c}$, so that

$$\ddot{Y}(y) = c \cdot e^{\ln \frac{y}{c} \cdot 2^{\frac{1}{1-\gamma}}} = c \cdot \left(\frac{y}{c}\right)^{2^{\frac{1}{1-\gamma}}} = y^{\ddot{\beta}} c^{1-\ddot{\beta}}.$$

□

The coefficient $\ddot{\beta}$ is the *income elasticity of protected income*. For $\gamma > 1$ the elasticity $\ddot{\beta}$ is thus between zero and one. This implies that $\frac{\ddot{Y}(y)}{y} = \left(\frac{c}{y}\right)^{1-\ddot{\beta}}$ is a decreasing function of equivalent income, and consequently the relative collateral damage $1 - \left(\frac{c}{y}\right)^{1-\ddot{\beta}}$ is an increasing function of equivalent income, as desired.

Note that $\ddot{\beta}$ is increasing in γ, that $\lim_{\gamma \to \infty} 2^{\frac{1}{1-\gamma}} = 1$, and that $\lim_{\gamma \downarrow 1} 2^{\frac{1}{1-\gamma}} = 0$. In words: a large γ corresponds to the situation where most of Ana's equivalent income is protected as Ben's income grows without bound whereas a small γ corresponds to a situation where only a baseline level c of Ana's income is protected as Ben's income grows without bound.

For these orderings, protected income is therefore a weighted geometric average of a normatively determined baseline income level, c, and the original income level, y, and where the weight on income is precisely the income elasticity $\ddot{\beta}$. Because this income elasticity is the same for high or low income

levels, we can identify these orderings as those having constant protected income elasticities and, when staying within this class, knowing the level $c > 0$ and the magnitude of the elasticity $\ddot{\beta}$ is sufficient to pin down the entire ordering, as one can recover the parameter γ through the expression $\ddot{\beta} = 2^{\frac{1}{1-\gamma}}$.

We show that constant protected income elasticities and a baseline protected income level indeed characterize this class of social welfare functions.

Proposition 7. *The subclass of social welfare functions generated by f (as defined in this section) is, up to an affine transform, characterized by the following property: There exist $\ddot{\beta}$, $\dddot{\beta} \in (0,1)$, $c > 0$* such that, for all $y \in (c, \infty)$,

$$\ddot{Y}(y) = y^{\ddot{\beta}} c^{1-\ddot{\beta}} \text{ and } \dddot{Y}(y) = y^{\dddot{\beta}} c^{1-\dddot{\beta}}.$$

In particular, necessarily there is $\gamma > 1$ such that $\ddot{\beta} = 2^{\frac{1}{1-\gamma}}$, $\dddot{\beta} = 3^{\frac{1}{1-\gamma}}$.

Proof. See Appendix 10.2. □

7 Unequal status quo

So far, we have taken equality as a starting point and examined what sacrifice could be imposed on Ana for the sake of one or two other individuals. It is interesting also to generalize the analysis and examine starting points containing some inequality.

Consider any starting point (y_1, y_2). How much of Ana's income is protected against a rise in Ben's income? This can be derived in straightforward manner from the previous notions. First, let us define the equally-distributed equivalent, relative to function f, of the unequal distribution (y_1, y_2) :

$$EE_f(y_1, y_2) := f^{-1}\left[\frac{1}{2}f(y_1) + \frac{1}{2}f(y_2)\right].$$

Retaining the same notation $\ddot{Y}$ for both an unequal and an equal status quo, it is clear that

$$\ddot{Y}(y_1, y_2) = \ddot{Y}(EE_f(y_1, y_2)),$$

since an egalitarian status quo that would yield the same social welfare as (y_1, y_2) is suitable as a benchmark for the social welfare level to keep constant while Ana's income is decreased for the sake of Ben's. Retaining the

convention that $f(+\infty) = 0$ for a bounded function, one also computes

$$\ddot{Y}(y_1, y_2) = f^{-1}(f(y_1) + f(y_2)).$$

It is worth observing that there is a correspondence between the properties of protected income (and collateral damage) and invariance properties of EE_f which also characterize the corresponding social welfare functions. Let $EE_\eta(y_1, y_2)$ denote the equally-distributed equivalent for a Kolm-Atkinson social preference with inequality aversion coefficient $\eta > 0$. Let $EE_\alpha(y_1, y_2)$ denote the equally-distributed equivalent for a Kolm-Pollak social preference with inequality aversion coefficient $\alpha > 0$. Last, given $c > 0$, $y_1 \geq c$ and $y_2 \geq c$, let $EE_\gamma(y_1, y_2, c)$ denote the equally-distributed equivalent for the social preference defined in Section 6 with inequality aversion coefficient $\gamma > 0$. We obtain the following:

Proposition 8. *For the Kolm-Atkinson class, $EE_\eta(y_1, y_2)$ is scale invariant: for every $\kappa > 0$,*

$$EE_\eta(\kappa y_1, \kappa y_2) = \kappa EE_\eta(y_1, y_2).$$

For the Kolm-Pollack class, $EE_\alpha(y_1, y_2)$ is translation invariant: for every $k > 0$,

$$EE_\alpha(y_1 + k, y_2 + k) = EE_\alpha(y_1, y_2) + k.$$

For the class introduced in Section 6, given $c > 0$, $y_1 \geq c$ and $y_2 \geq c$, $EE_\gamma(y_1, y_2, c)$ is compound invariant: for every $\rho > 0$,

$$EE_\gamma(y_1^\rho, y_2^\rho, c^\rho) = EE_\gamma(y_1, y_2, c)^\rho.$$

Proof. This is known for the first two classes, and we only prove the third case.

Let $\gamma \neq 1$. Then $EE = EE_\gamma(y_1, y_2, c)$ satisfies

$$\frac{1}{2}\frac{\left[\ln\left(\frac{y_1}{c}\right)\right]^{1-\gamma}}{1-\gamma} + \frac{1}{2}\frac{\left[\ln\left(\frac{y_2}{c}\right)\right]^{1-\gamma}}{1-\gamma} = \frac{\left[\ln\left(\frac{EE}{c}\right)\right]^{1-\gamma}}{1-\gamma}.$$

Let $\rho > 0$. Then $EE^* = EE_\gamma(y_1^\rho, y_2^\rho, c^\rho)$satisfies

$$\frac{1}{2}\frac{\left[\ln\left(\frac{y_1}{c}\right)^\rho\right]^{1-\gamma}}{1-\gamma} + \frac{1}{2}\frac{\left[\ln\left(\frac{y_2}{c}\right)^\rho\right]^{1-\gamma}}{1-\gamma} = \frac{\left[\ln\left(\frac{EE^*}{c^\rho}\right)\right]^{1-\gamma}}{1-\gamma},$$

that is,

$$\frac{1}{2}\frac{\left[\rho\ln\left(\frac{y_1}{c}\right)\right]^{1-\gamma}}{1-\gamma} + \frac{1}{2}\frac{\left[\rho\ln\left(\frac{y_2}{c}\right)\right]^{1-\gamma}}{1-\gamma} = \frac{\left[\ln\left(\frac{EE^*}{c^\rho}\right)\right]^{1-\gamma}}{1-\gamma},$$

$$\frac{1}{2}\frac{\rho^{1-\gamma}\left[\ln\left(\frac{y_1}{c}\right)\right]^{1-\gamma}}{1-\gamma}+\frac{1}{2}\frac{\rho^{1-\gamma}\left[\ln\left(\frac{y_2}{c}\right)\right]^{1-\gamma}}{1-\gamma}=\frac{\left[\ln\left(\frac{EE^*}{c^\rho}\right)\right]^{1-\gamma}}{1-\gamma}.$$

$$\rho^{1-\gamma}\frac{\left[\ln\left(\frac{EE}{c}\right)\right]^{1-\gamma}}{1-\gamma}=\frac{\left[\ln\left(\frac{EE^*}{c^\rho}\right)\right]^{1-\gamma}}{1-\gamma}$$

$$\frac{\ln\left(\left(\frac{EE}{c}\right)^\rho\right)^{1-\gamma}}{1-\gamma}=\frac{\left[\ln\left(\frac{EE^*}{c^\rho}\right)\right]^{1-\gamma}}{1-\gamma}$$

from where it follows that $EE^\rho = EE^*$.

Now let $\gamma = 1$. Then $EE = EE_1(y_1, y_2, c)$ satisfies

$$\frac{1}{2}\ln\ln\frac{y_1}{c}+\frac{1}{2}\ln\ln\frac{y_2}{c}=\ln\ln\left(\frac{EE}{c}\right).$$

Fix $\rho > 0$. Then $EE^* = EE_1(y_1^\rho, y_2^\rho, c^\rho)$satisfies

$$\frac{1}{2}\ln\ln\left(\frac{y_1}{c}\right)^\rho+\frac{1}{2}\ln\ln\left(\frac{y_2}{c}\right)^\rho=\ln\ln\left(\frac{EE^*}{c^\rho}\right),$$

that is,

$$\frac{1}{2}\ln\left(\rho\ln\frac{y_1}{c}\right)+\frac{1}{2}\ln\left(\rho\ln\frac{y_2}{c}\right)=\ln\ln\left(\frac{EE^*}{c^\rho}\right),$$

$$\frac{1}{2}\left[\ln\rho+\ln\ln\frac{y_1}{c}\right]+\frac{1}{2}\left[\ln\rho+\ln\ln\frac{y_2}{c}\right]=\ln\ln\left(\frac{EE^*}{c^\rho}\right),$$

$$\ln\rho+\frac{1}{2}\ln\ln\frac{y_1}{c}+\frac{1}{2}\ln\ln\frac{y_2}{c}=\ln\ln\left(\frac{EE^*}{c^\rho}\right),$$

$$\ln\rho+\ln\ln\left(\frac{EE}{c}\right)=\ln\ln\left(\frac{EE^*}{c^\rho}\right),$$

$$\ln\left(\rho\ln\left(\frac{EE}{c}\right)\right)=\ln\ln\left(\frac{EE^*}{c^\rho}\right),$$

$$\ln\ln\left(\frac{EE}{c}\right)^\rho=\ln\ln\left(\frac{EE^*}{c^\rho}\right),$$

from where it follows that $EE^\rho = EE^*$, which is what we wanted to show. □

8 Conclusions

With the contribution of this paper, the menu of choices for how to elicit ethical preferences about inequality aversion is enriched with a framing that is quite different from the leaky-bucket experiment usually employed, explicitly or implicitly, in surveys and experiments. Giving sacrificed individuals a protection against the collateral damage of pursuing the improvement in the lot of other people seems an attractive principle, and it can serve to calibrate the coefficient of inequality aversion.

Moreover, we conjecture that this thought experiment tends to justify higher values of inequality aversion than the leaky-bucket experiment. Two cognitive mechanisms may induce this phenomenon. First, the leaky-bucket thought experiment involves a loss to society, as part of the transfer is wasted, and loss aversion is likely to kick in and generate a jaundiced view of the transfer. Second, the usual application involving a discrepancy between finite transfers and a numerical calibration of the coefficient that is only justified for infinitesimal transfers biases the analysis in the direction of a smaller coefficient (e.g., the proportional transfer principle requires a coefficient of 2 but only a coefficient of 1 for infinitesimal transfers, and it is wrong to infer from the respondents' acceptance of proportional transfers that their coefficient is 1).

In the thought experiment proposed here, loss aversion pushes in the direction of greater inequality aversion, since the loss that is imposed on the sacrificed individual looms large in the ethical evaluation. We therefore propose that empirical estimations of inequality aversion in the population should no longer be confined to leaky-bucket types of experiments (and certainly not confuse infinitesimal and large transfers) and should instead, or at least additionally, involve a "protected level" type of experiment in order to obtain another relevant estimate.

A separate contribution of the paper is a characterization and representation of social preferences using parametrizations with straightforward interpretations. To wit: the Kolm-Pollak social preferences can be described by noting that they exhibit *constant difference protected income* (CDPI), and we write $f(y) = -2^{-\frac{y}{L}}$ where $L > 0$ identifies the largest collateral damage that is acceptable to the evaluator in a two-person evaluation for all income levels, and therefore $\ddot{Y}(y) = y - L$ for all $y > 0$. In turn, the Kolm-Atkinson social preferences that display positive protection can be described by noting that they exhibit *constant relative protected income* (CRPI), and we write

$f(y) = -y^{\frac{1}{\log_2 \lambda}}$ where $\lambda \in (0,1)$ identifies the smallest fraction of income that will be protected by the evaluator in a two-person evaluation for all income levels, and therefore $\ddot{Y}(y) = \lambda y$ for all $y > 0$. Last, we can say that the sub-class of social preferences we proposed in Section 6 that displays positive protection exhibits *constant protected income elasticities* (CPIE), and we write $f(y) = -\left(\ln \frac{y}{c}\right)^{\frac{1}{\log_2 \beta}}$ for $y > c$, where $\beta \in (0,1)$ is the income elasticity of protected income for all income levels in a two-person evaluation, c is a normatively determined lower bound on protected income for all $\beta \in (0,1)$, and $\ddot{Y}(y) = y^{\beta} c^{1-\beta}$ for all $y > c$. Having representations of social preferences that are easy to motivate and interpret, as we have aimed to provide in this paper and its companion website, https://osf.io/tnu2q/, can lower the barriers to adoption of these social welfare tools by applied economists. This would be a welcome development to anyone interested in fostering the inequality sensitive evaluation of economic and social policies.

9 References

Aboudi, R., & Thon, D. (2003). Transfer principles and relative inequality aversion a majorization approach. Mathematical social sciences, 45(3), 299-311.

Aboudi, R., & Thon, D. (2006). Refinements of Muirhead's Lemma and income inequality. Mathematical social sciences, 51(2), 201-216.

Atkinson, A. B. (1970), "On the Measurement of Inequality," Journal of Economic Theory, 2, 244–263.

Bosmans, K. (2014). Distribution-sensitivity of rank-dependent poverty measures. Journal of Mathematical Economics, 51, 69-76.

Chakravarty, S. R. (2009). Inequality, polarization and poverty. Advances in distributional analysis. New York.

Cowell, F. A. (2000). Measurement of inequality. *Handbook of income distribution*, vol. 1, 87–166.

Dubois, M. (2016). A note on the normative content of the Atkinson inequality aversion parameter. Economics Bulletin.

Dutta, B. (2002). Inequality, poverty and welfare. Handbook of social choice and welfare, 1, 597-633.

Fleurbaey M., P. Michel (2001). "Transfer principles and inequality aversion, with an application to optimal growth," *Mathematical Social Sciences* 42: 1–11.

Gaertner, W., & Schokkaert, E. (2012). Empirical social choice: questionnaire-experimental studies on distributive justice. Cambridge University Press.

Hurley, J., Mentzakis, E., & Walli-Attaei, M. (2020). Inequality aversion in income, health, and income-related health. Journal of health economics, 70, 102276.

Kolm, S. C. (1976). Unequal inequalities. I. Journal of economic Theory, 12(3), 416-442.

Okun, A. M. (2015). Equality and efficiency: The big tradeoff. Brookings Institution Press.

Thon D., S.W. Wallace (2004). "Dalton transfers, inequality and altruism," Social Choice and Welfare 22: 447-465.

Tossavainen T., P. Haukkanen (2007). "The Schröder Equation and Some Elementary Functions", *Int. Journal of Math. Analysis* 1(3): 101–106.

Venmans, F., & Groom, B. (2021). Social discounting, inequality aversion, and the environment. Journal of Environmental Economics and Management, 109, 102479.

10 Appendix

10.1 Proof of Proposition 5

Proof. From Prop. 4 we know that $f(y) = -e^{g(y)}$ for a function g such that

$$g((k+1)\Delta + a) - g(k\Delta + a) = -\ln 2$$

for all $k \in \mathbb{Z}_+$,$a \in [0, \Delta]$. By a similar reasoning, the function g must also satisfy

$$g((k+1)\Omega + a) - g(k\Omega + a) = -\ln 3$$

for all $k \in \mathbb{Z}_+$,$a \in [0, \Omega]$. In particular, for all $k \in \mathbb{Z}_+$,

$$g(k\Delta) - g(0) = -k \ln 2$$

$$g(k\Omega) - g(0) = -k \ln 3.$$

Given the quasi-periodic behavior of g, there exist $b < 0 < c$ such that for all $x \in \mathbb{R}$,

$$b < g(x) - g(0) + x\frac{\ln 2}{\Delta} < c$$

$$b < g(x) - g(0) + x\frac{\ln 3}{\Omega} < c.$$

This is possible only if

$$\frac{\ln 2}{\Delta} = \frac{\ln 3}{\Omega}.$$

Let α be this common ratio. This proves that $\Delta = \frac{\ln 2}{\alpha}$ and $\Omega = \frac{\ln 3}{\alpha}$ for some $\alpha > 0$.

Without loss of generality, we can let $f(+\infty) = 0$ and $f(0) = -1$. Then

$$\begin{cases} f(y-\Delta) = 2f(y) \\ f(y-\Omega) = 3f(y). \end{cases}$$

Letting $h = f \circ ln$, a change of variables $y = \ln x$ yields

$$\begin{cases} h\left(2^{-\frac{1}{\alpha}}x\right) = 2h(x) \\ h\left(3^{-\frac{1}{\alpha}}x\right) = 3h(x). \end{cases}$$

The rest of the argument follows the proof of the bottom half of Prop. 3, for $\eta = 1+\alpha$, from where we obtain that $h(x) = h(1)\,x^{-\alpha}$ for all $x \in (0, +\infty)$. Then $f(y) = h(e^y) = h(e^0)\,e^{-\alpha y} = f(0)\,e^{-\alpha y} = -e^{-\alpha y}$ for all $y \in \mathbb{R}$. □

10.2 Proof of Proposition 7

Proof. The proof mimics previous proofs. First, one shows that $\ddot{\beta} = 2^{\frac{1}{1-\gamma}}$, $\dddot{\beta} = 3^{\frac{1}{1-\gamma}}$. To do this, we note that, by definition, $\frac{1}{2}f\left(\ddot{Y}(y)\right) + \frac{1}{2}f(\infty) = f(y)$. Notice that $\ddot{Y}(y) \in (0, y)$ and since f is monotone increasing this implies that f is bounded from above. Without loss of generality, we can let $f(+\infty) = 0$ and $f(c \cdot e) = \frac{1}{1-\gamma}$. Then $f\left(y^{\ddot{\beta}}c^{1-\ddot{\beta}}\right) = 2f(y)$. Similarly, we know that $\frac{1}{3}f\left(\dddot{Y}(y)\right) + \frac{1}{3}f(\infty) + \frac{1}{3}f(\infty) = f(y)$ and so $f\left(y^{\dddot{\beta}}c^{1-\dddot{\beta}}\right) = 3f(y)$. Let $z = \ln y - \ln c$ and $h(z) = f(c \cdot e^z)$. Then the equations $f\left(y^{\ddot{\beta}}c^{1-\ddot{\beta}}\right) = 2f(y)$ and $f\left(y^{\dddot{\beta}}c^{1-\dddot{\beta}}\right) = 3f(y)$ respectively become

$$\begin{cases} h\left(\ddot{\beta}z\right) = 2g(z) \\ h\left(\dddot{\beta}z\right) = 3g(z). \end{cases}$$

Now let $g = \ln(-h \circ e)$. A second change of variables $x = \ln z$ yields

$$\begin{cases} \ln 2 + g(x) = g\left(\ln \ddot{\beta} + x\right) \\ \ln 3 + g(x) = g\left(\ln \dddot{\beta} + x\right). \end{cases}$$

We obtain that, for all $k \in \mathbb{Z}_+$,

$$g\left(k \ln \ddot{\beta}\right) - g(0) = k \ln 2$$

$$g\left(k \ln \dddot{\beta}\right) - g(0) = k \ln 3$$

which, as in the proof of Prop. 3, is only possible if

$$\frac{\ln 2}{\ln \ddot{\beta}} = \frac{\ln 3}{\ln \dddot{\beta}}.$$

We now let $1-\gamma$ be this common (negative) ratio. This proves that $\ddot{\beta} = 2^{\frac{1}{1-\gamma}}$ and $\dddot{\beta} = 3^{\frac{1}{1-\gamma}}$ for some $\gamma > 1$.

The rest of the argument is as in the second half of the proof of Prop. 3, from where it follows that $h(z) = h(1) z^{1-\gamma}$ for all $z \in (0, \infty)$. Since $h(1) = f(c \cdot e) = \frac{1}{1-\gamma}$, we obtain that $f(y) = g(\ln y - \ln c) = \frac{\left(\ln \frac{y}{c}\right)^{1-\gamma}}{1-\gamma}$ for all $y \in (c, \infty)$, which is what we wanted to show. □