Department of Economics

DISCUSSION PAPER SERIES IN ECONOMICS

DP No. 12

# Individual Sense of Justice and Harsanyi's Impartial Observer

June 2019

Abhinash Borah

https://www.ashoka.edu.in/ecodp

# Individual Sense of Justice and Harsanyi's Impartial Observer

Abhinash Borah*

June 7, 2019

**Abstract**

We revisit, within Harsanyi's impartial observer setting, the question of foundations underlying procedural fairness concerns in welfare judgments. In our setup—that of allocating an *indivisible* good using a lottery—such concerns, presumably, matter. We draw from the social preferences literature and relax a typical assumption made while addressing this question, namely, that individuals in society do not care about procedural fairness and such concerns arise exclusively at a societal level, which are captured by non-linear social welfare functions (SWFs). In our model, individual attitudes towards procedural fairness are identified and factored into welfare judgments. Specifically, we provide an axiomatic basis within Harsanyi's (1955) framework to represent procedural fairness sensitive individual preferences by the representation in Karni and Safra (2002). We then show, in terms of underlying axioms, how such individual assessments incorporating both risk and procedural fairness attitudes can be aggregated by means of utilitarian and generalized utilitarian SWFs.

**Keywords**: Procedural fairness, Harsanyi's impartial observer, Karni-Safra ("individual sense of justice") preferences, social preferences under risk, utilitarianism, generalized utilitarianism

**JEL Classification**: D63, D71, D81

*Department of Economics, Ashoka University. Email: abhinash.borah@ashoka.edu.in. Address: Plot No. 2, Rajiv Gandhi Education City, National Capital Region, P.O. Rai, Sonepat, Haryana - 131029, India. I am grateful for comments made by participants attending the 2018 RUD Conference at Heidelberg University. I have also gained much from several illuminating conversations on the subject of procedural fairness with Andy Postlewaite and Alavaro Sandroni.

# 1 Introduction

Consider the classic distributional problem of deciding who amongst a set of potential claimants should be allocated an indivisible good. In this setting, the ex-post allocation is necessarily unfair as only one person can receive the good. As a way of compensating for this, an emphasis is often placed on allocating the good in a way that is procedurally fair. One simple and popular way of achieving such procedural fairness is by using a lottery to allocate the good. For instance, consider the problem of allocating one available kidney amongst two equally deserving patients, Tom and Bob, who are both in need of a kidney transplant. In a situation like this, it is not uncommon to use a fair lottery to determine the allocation on the ground that it equalizes the ex ante opportunities of the two individuals and, hence, is procedurally fair.[1] This paper revisits, within Harsanyi's impartial observer setting [Harsanyi (1955), Harsanyi (1953)], the question of the foundations underlying such procedural fairness sensitive welfare judgments in a set-up where an indivisible good is allocated by means of a lottery. Specifically, we relate this question to the literature on social preferences under risk and explore the possibility of drawing on individual subjective attitudes towards procedural fairness in forming such judgments.

Harsanyi's impartial observer setting is premised on the observation that one way of arriving at normatively acceptable welfare judgments is to take the perspective of an impartial observer who is behind a *hypothetical* "veil of ignorance" and faces risk about his identity in society. Such risk, goes the argument, would force him to weigh the well-being of all members of society under alternative social states (e.g., allocations, as in our case) and make ethically acceptable interpersonal comparisons while forming welfare judgments. To that end, Harsanyi assumed that each individual is characterized by two sets of preferences. First, he has his standard *subjective preferences* that capture what he actually prefers or chooses. Second, he has his *ethical preferences* that capture his welfare judgments made from the perspective of an impartial observer. Harsanyi showed that if both subjective as well as ethical preferences satisfy the independence axiom and the acceptance

[1] Examples of lotteries being used to distribute scarce resources can be found in the allocation of public housing, admission to educational institutions, athletic drafts (e.g., the National Basketball Association), US green cards, (avoiding) military drafts and, indeed, medical resources such as kidney transplants.

principle holds,[2] then any impartial observer's ethical preferences (welfare judgments) have to conform to the logic of a utilitarian social welfare function (SWF). That is, while comparing lotteries over social states, it is as if, his assessments of these lotteries are based on a weighted average of individual expected utilities under them.

A utilitarian SWF is sensitive only to the sum total of individual utilities and not to the distribution of these utilities. Hence, it fails to discriminate between allocation procedures on the basis of the fairness of this distribution and, consequently, on the basis of whether the ex ante opportunities that different individuals have are fair or not. Because of this, it cannot accommodate welfare judgments like that of using a lottery to allocate an indivisible good as in the kidney allocation example above.[3] Given that such welfare judgments accommodating procedural fairness concerns are both intuitively appealing and find resonance in how many real-world allocation problems are resolved, the literature has looked at ways to overcome this implication of Harsanyi's utilitarianism. A leading research question in this area, therefore, has been about proposing foundations underlying such welfare judgments that are sensitive to procedural fairness concerns.

The popular approach that the literature appears to have converged to in terms of addressing this issue is to employ non-linear or non-utilitarian SWFs, e.g., SWFs that are concave in individual expected utilities. Such SWFs are sensitive to the ex ante distribution of individual expected utilities and therefore can accommodate

[2]The acceptance principle requires that when an impartial observer imagines himself to be a particular individual, he should adopt that individual's preferences.

[3]This is an observation that dates back to Diamond (1967). To see this, let $(1,0)$ and $(0,1)$, respectively, denote the allocations under which Tom and Bob receive the kidney. Let $u_i(1,0)$ and $u_i(0,1)$, respectively, denote the utility of individual $i$ = Tom ($T$), Bob ($B$) under these two allocations. It seems reasonable to assume that $u_T(1,0) > u_T(0,1)$ and $u_B(1,0) < u_B(0,1)$. Further, from the perspective of an impartial observer's interpersonal comparisons, he is presumably indifferent between being Tom under the allocation $(1,0)$ and being Bob under the allocation $(0,1)$; and between being Tom under the allocation $(0,1)$ and Bob under the allocation $(1,0)$. This implies that $u_T(1,0) = u_B(0,1) =: \overline{u} > \underline{u} := u_T(0,1) = u_B(1,0)$. Assume that an impartial observer faces an equal chance of being Tom or Bob. Then, under Harsanyi's utilitarianism, his assessment of the lottery $[(1,0), .5; (0,1), .5]$, which gives Tom and Bob an equal chance of receiving the kidney, is given by a simple average of Tom's and Bob's expected utilities under this lottery, i.e., $.5(.5u_T(1,0) + .5u_T(0,1)) + .5(.5u_B(1,0) + .5u_B(0,1)) = .5(.5\overline{u} + .5\underline{u}) + .5(.5\underline{u} + .5\overline{u}) = .5\overline{u} + .5\underline{u}$. At the same time, his assessment of the degenerate lottery under which, say, Tom receives the kidney for sure is given by $.5u_T(1,0) + .5u_B(1,0) = .5\overline{u} + .5\underline{u}$. Given that these assessments are the same, he is indifferent between the two lotteries.

procedural fairness concerns.[4] In terms of Harsanyi's impartial observer set-up, this alternative approach of using non-linear SWFs to accommodate procedural fairness concerns translates to the following. In Harsanyi, both subjective preferences as well as ethical impartial observer preferences are of the expected utility type. Under this alternative approach, whereas subjective preferences are still of the expected utility type, preferences of an impartial observer need not be. It is this relaxation of Bayesian rationality at the level of an impartial observer's preferences that provides the flexibility to make welfare judgments on the basis of a non-linear SWF. What this means is that, under this approach, the actual subjective preferences of individuals in society do not show any concern for procedural fairness as they are of the expected utility type. Rather, the exclusive source of such concerns in welfare judgments is restricted to impartial observer preferences. In other words, the possibility that this important consideration in impartial observer preferences and welfare may have a basis in the subjective preferences of individuals in society is a priori ruled out. This observation is our point of departure in this paper.

We draw on the findings of an emerging literature on social preferences under risk and consider the possibility that subjective preferences of individuals may indeed exhibit a concern for procedural fairness, especially in situations where unfair ex-post allocations are inevitable. If that is the case, then two observations are worth noting. First, at a descriptive level, the aforementioned modelling approach is inadequate to accommodate such individuals as their subjective preferences will not be of the expected utility type. For instance, in the context of the kidney allocation example mentioned above, suppose Tom strictly prefers the allocation in which he gets the kidney to the allocation in which Bob gets it. At the same time, suppose he also strictly prefers the lottery that gives Bob a 10% chance of

[4]Refer to Epstein and Segal (1992) and Grant, Kajii, Polak, and Safra (2010), among others. For instance, in the *generalized utilitarian* formulation of Grant, Kajii, Polak, and Safra (2010), in order to accommodate procedural fairness concerns, an impartial observer transforms the individual expected utilities using a strictly concave function. Continuing with the kidney example and the notation from footnote 3, under generalized utilitarianism, such an impartial observer's assessment of the lottery $[(1,0), .5; (0,1), .5]$ is given by $.5\phi(.5u_T(1,0) + .5u_T(0,1)) + .5\phi(.5u_B(1,0) + .5u_B(0,1)) = .5\phi(.5\overline{u} + .5\underline{u}) + .5\phi(.5\underline{u} + .5\overline{u}) = \phi(.5\overline{u} + .5\underline{u})$, where $\phi : \mathbb{R} \to \mathbb{R}$ is a strictly increasing and strictly concave function. On the other hand, his assessment of the degenerate lottery under which Tom receives the kidney for sure is given by $.5\phi(u_T(1,0)) + .5\phi(u_B(1,0)) = .5\phi(\overline{u}) + .5\phi(\underline{u})$. Clearly, for a strictly concave $\phi$, $\phi(.5\overline{u} + .5\underline{u}) > .5\phi(\overline{u}) + .5\phi(\underline{u})$. That is, under generalized utilitarianism, an impartial observer strictly prefers the coin toss to determine who receives the kidney.

getting the kidney and him a 90% chance to the degenerate lottery under which he gets the kidney for sure.[5] Such preferences can be directly attributed to the fact that sharing ex ante chances or opportunities with Bob is a way for Tom to ensure a degree of procedural fairness in a situation where outcome fairness is impossible. However, note that these preferences violate the independence axiom and, hence, cannot have an expected utility representation. Second, when it comes to the question of providing foundations for procedural fairness concerns in welfare judgments, one ought to account for and draw on the information contained in individual subjective preferences regarding sensitivity towards such concerns. That is, if individuals in society themselves care about procedural fairness, then it seems only reasonable that these attitudes be identified and factored into welfare judgments. Impartial observer preferences need not necessarily be the exclusive source of such concerns in welfare judgments, independent of what individual attitudes with respect to such concerns are. That is the task that we formally undertake in this paper.

Our starting point, like in Harsanyi's original formulation, is to consider individuals who have two sets of preferences: subjective preferences and ethical (impartial observer) preferences. The economic problem at hand is that of allocating an indivisible good by means of a lottery. The key innovation in our model as compared to the existing literature in this area is that subjective preferences of individuals in this context may show a concern for procedural fairness, e.g., like that of Tom above. If this is the case, then these preferences are not of the expected utility type. So, the first main task of the paper is to suggest a non-expected utility representation for such preferences that can accommodate concerns for procedural fairness. Here, we draw inspiration from Karni and Safra's influential paper, "Individual Sense of Justice: A Utility Representation" (Karni and Safra, 2002). Their paper proposes and provides an axiomatic foundation for a representation of procedural fairness sensitive individual preferences in an economic environment

[5]Experimental evidence suggests that such preferences are plausible. For instance, consider the two player probabilistic dictator game. In such a game, a decision maker (the "dictator") is endowed with a fixed amount of money. He is not allowed to share the money with the other individual, but he is given the option, if he so chooses, to share *chances* of getting the money with him, i.e., he can assign the other individual any probability of getting the entire amount while retaining the amount himself with complementary probability. Experimental evidence [e.g., Krawczyk and LeLec (2010), Brock, Lange, and Ozbay (2013)] suggests that a significant portion of decision makers do give the other individual a positive probability (on average of about 0.1 in the experiments) of getting the money.

identical to ours, i.e., of using a lottery to allocate an indivisible good. Their concern in that paper is exclusively with individual preferences and not welfare. Their primitive preference relations are different from ours and their axiomatization is not directly applicable to our set-up. The first key result of this paper shows though that it is possible, within Harsanyi's framework, to accommodate individual subjective preferences that are sensitive to procedural fairness concerns in the Karni–Safra sense. Specifically, we provide an axiomatic foundation within this framework for such preferences to have a Karni–Safra representation. This representation of subjective preferences allows us, within the Harsanyi set-up, to identify and provide a sharp separation between an individual's attitudes towards risk and procedural fairness in assessing lotteries determining the allocation of the indivisible good.

Thereafter, we extend the axiomatic framework to show how the Karni-Safra assessments of individuals who do care about procedural fairness and the expected utility assessments of individuals with "standard" preferences who don't can together be faithfully incorporated into a SWF. In keeping with the Harsanyi tradition, our view of any such SWF is a subjectivist one and we view it as a representation of an impartial observer's ethical preferences. We show that we can do this exercise for both utilitarian as well as generalized utilitarian SWFs and we identify axiomatic bases for both. As such, by demonstrating how individual attitudes towards procedural fairness may be identified and incorporated in welfare judgments, our results provide a foundation for procedural fairness concerns in such judgments in terms of actual preference attitudes of individuals in society. This means that in our analysis, procedural fairness sensitive welfare judgments need not be exclusively based on a paternalistic concern but rather can be based on individual values as well. Therefore, it is closer in spirit to how economists view the philosophical underpinnings of welfare economics.

We are not the first to suggest that concerns for procedural fairness be accommodated in social welfare judgments based on an "all-inclusive" notion of individual utilities that can capture such concerns. Such a suggestion finds resonance in Broome (1984, 1991), Karni (1996) and Trautmann (2010), among others. We add to this literature by offering an axiomatization that clarifies the exact manner in which, starting from individual preferences, we can derive all-inclusive, procedure-sensitive individual utilities and accommodate them in both utilitarian and generalized utilitarian SWFs. In other words, the claim that social welfare

judgments can draw on all-inclusive, procedure-sensitive individual utilities is not an assumption in our model, but rather, it follows from our axioms.

In this paper, we also draw on the literature on social preferences. The first generation of social preference models (e.g., Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Charness and Rabin (2002)) were proposed for risk-free environments. The literature soon discovered that that these models cannot be readily extended to environments of risk using standard approaches like expected utility or the available non-expected utility theories. In simple terms, this is because these theories of decision making under risk cannot accommodate a "preference for randomization" that may arise, as we have seen, owing to procedural fairness concerns.[6] Hence, the more recent attempts in the literature have been to develop models of social preferences under risk that can accommodate procedural fairness concerns. Such attempts have been made by Fudenberg and Levine (2012) and Saito (2013), amongst others. The paper we draw on most here, Karni and Safra (2002), is an early precursor to this line of research.

Finally, when it comes to the issue of aggregating individual assessments using utilitarian and generalized utilitarian SWFs and the axiomatic foundations for such an exercise, we have drawn from Grant, Kajii, Polak, and Safra (2010) and Karni and Safra (2000).

The rest of the paper is organized as follows. Section 2 sets up the framework. Section 3 shows how, for individuals who care about procedural fairness, attitudes towards it can be separated from attitudes towards risk. In this section, we formally define and axiomatize a Karni-Safra representation of subjective preferences. Finally, in Section 4, we show, based on underlying axioms, how individual attitudes towards both risk and procedural fairness can be aggregated and accommodated within both utilitarian and generalized utilitarian SWFs. Proofs of results are provided in the Appendix.

[6] In slightly more technical terms, all standard models of decision making under risk satisfy the property of stochastic dominance. On the other hand, social preferences under risk may often violate this property. For instance, Tom's suggested behavior above, which essentially replicates the choices of many decision makers in the probabilistic dictator game, violates stochastic dominance and, hence, cannot be accommodated by these standard models.

# 2 Framework

## 2.1 Preliminaries

We consider a society comprising of a finite number of individuals, with $I = \{1, \ldots, \overline{I}\}$ denoting the set of individuals, $\overline{I} \geq 2$, and $i$, $j$ its typical elements. There is one unit of an *indivisible good* that must be allocated to one of the $\overline{I}$ individuals. Accordingly, the set of allocations for this society is given by:

$$X = \{x = (x(1), \ldots, x(\overline{I})) \in \mathbb{R}^{\overline{I}} : x(i) \in \{0, 1\} \text{ and } \textstyle\sum_{i=1}^{\overline{I}} x(i) = 1\},$$

with $x(i)$ denoting the number of units of the good that individual $i \in I$ receives under the allocation $x \in X$. We denote the set of simple lotteries on the sets $I$ and $X$ by $\Delta(I)$ and $\Delta(X)$, respectively. We refer to elements of $\Delta(I)$ as *identity lotteries* and denote a typical element from this set by $z$, with $z(i)$ denoting the probability assigned by $z$ to individual $i \in I$. On the other hand, we refer to elements of $\Delta(X)$ as *outcome lotteries* and denote a typical element from this set by $l$, with $l(x)$ denoting the probability assigned by $l$ to the allocation $x \in X$.

Besides the standard interpretation of an outcome lottery as specifying the risk pertaining to the final allocation (*allocation risk*, for short), in our analysis, it has the additional interpretation of being an *allocation procedure* through which society solves its distributional problem of allocating the one unit of the indivisible good among the $\overline{I}$ individuals. When viewed from the perspective of being an allocation procedure, among other things, an outcome lottery can be identified with the opportunity that different individuals in society have of receiving the good and the fairness of these opportunities, i.e., it can be identified with a notion of procedural fairness. In general, we should expect individual attitudes towards the allocation risk and that towards the allocation procedure to be distinct considerations influencing the assessment of outcome lotteries, and our goal here is to incorporate this distinction into social welfare judgments.

We assume that $\Delta(I)$ and $\Delta(X)$ are endowed with the Euclidean topology. Further, we refer to elements of the set $\Delta(I) \times \Delta(X)$ as *identity-outcome lotteries*. We assume that $\Delta(I) \times \Delta(X)$ is endowed with the product topology. When considering an identity-outcome lottery $(z, l) \in \Delta(I) \times \Delta(X)$, we assume that the identity

lottery $z$ and the outcome lottery $l$ are independently distributed. Given this independence assumption and denoting the set of simple lotteries on the set $I \times X$ by $\Delta(I \times X)$, we can equivalently view any identity-outcome lottery $(z, l)$ in terms of the product measure in $\Delta(I \times X)$ derived from $z$ and $l$, which we denote by $(z, l)^*$. That is, $(z, l)^* \in \Delta(I \times X)$ is the lottery that assigns the identity-allocation pair $(i, x) \in I \times X$ the probability $z(i) \times l(x)$. We define a convex combination of lotteries in the set $\Delta(X)$ or $\Delta(I)$, say, $\alpha l + (1-\alpha)l'$ or $\alpha z + (1-\alpha)z'$, $\alpha \in [0, 1]$, in the standard way. We denote any degenerate lottery by placing the outcome to which the lottery assigns unit probability within [.]-brackets. For instance, $[i] \in \Delta(I)$ and $[x] \in \Delta(X)$ denote degenerate lotteries that assign unit probability to $i \in I$ and $x \in X$, respectively. Non-degenerate lotteries are often denoted by explicitly listing out the possible realizations along with their respective probabilities in the standard way. For instance, $[x_1, \alpha_1; \ldots; x_M, \alpha_M]$ denotes the outcome lottery under which the allocation $x_m$ is realized with probability $\alpha_m$, $m = 1, \ldots, M$.

## 2.2 Preferences

We follow Harsanyi (1955) and assume that each individual $i \in I$ has two sets of preferences. First, he has a *subjective preference relation* $\succcurlyeq_i \subseteq \Delta(X) \times \Delta(X)$ over the set of outcome lotteries that expresses what he "actually prefers, whether on the basis of his personal interests or any other basis."[7] That is, $\succcurlyeq_i$ has the standard interpretation of a revealed preference relation over outcome lotteries. In contrast to Harsanyi's original formulation though, in our setting, preference judgments under $\succcurlyeq_i$ may reflect not just attitudes towards the allocation risk but also towards the allocation procedure, specifically concerns about procedural fairness. Second, he has an *ethical preference relation* $\succcurlyeq_i^* \subseteq (\Delta(I) \times \Delta(X)) \times (\Delta(I) \times \Delta(X))$ over the set of identity-outcome lotteries that expresses what he "prefers (or, rather, would prefer) on the basis of impersonal social considerations alone."[8] That is, this preference relation expresses his assessment of outcome lotteries when, instead of looking at them from his personal viewpoint, he does so from the perspective of an *impartial observer* in society whose assessment incorporates the well-being of all members of society in an impersonal way. The way impartiality and impersonality is incorporated in $\succcurlyeq_i^*$ is by maintaining that when individual $i$ imagines himself as

[7] Pg. 315, Harsanyi (1955)
[8] Op. cit.

an impartial observer and faces an identity-outcome lottery $(z, l)$, he is uncertain not only about which allocation will result but also about which person's identity he will assume in the given society, with the former uncertainty resolved according to the outcome lottery $l$, the latter according to the identity lottery $z$, and the two lotteries being independently distributed. It is worth pointing out that whereas outcome lotteries represent real risks, identity lotteries represent only hypothetical ones. In the process of ranking identity-outcome lotteries, by being required to "face" such hypothetical risks pertaining to his identity, he is forced to weigh the well-being of different individuals under alternative outcome lotteries, i.e., make interpersonal comparisons in an impartial and impersonal way. Accordingly, the preference relation $\succcurlyeq_i^*$ may be interpreted as capturing $i$'s social welfare judgments made from the perspective of an impartial observer.

We assume that, for each $i \in I$, $\succcurlyeq_i$ is complete and transitive. We denote the symmetric and asymmetric components of $\succcurlyeq_i$ by $\sim_i$ and $\succ_i$, respectively. In addition, to keep the problem meaningful, we assume that there exists at least some $i \in I$ for whom $\succ_i \neq \emptyset$. Similarly, for each $i \in I$, $\succcurlyeq_i^*$ is also assumed to be complete and transitive. We denote the symmetric and asymmetric components of $\succcurlyeq_i^*$ by $\sim_i^*$ and $\succ_i^*$, respectively. We further assume that this preference relation is continuous. That is, for any $(z', l') \in \Delta(I) \times \Delta(X)$, the sets $\{(z, l) \in \Delta(I) \times \Delta(X) : (z, l) \succ_i^* (z', l')\}$ and $\{(z, l) \in \Delta(I) \times \Delta(X) : (z', l') \succ_i^* (z, l)\}$ are open in $\Delta(I) \times \Delta(X)$.

**Remark 2.1.** There are two approaches that one may take when thinking about the identity of an impartial observer. The first is to think of an impartial observer as someone different from the members of society. The second, which is in line with Harsanyi's own interpretation, is to think of an impartial observer as a member of this society. We take the second approach here and assume that each individual, in addition to his subjective personal preferences, is able to make impersonal ethical judgements from the perspective of an impartial observer.

**Remark 2.2.** It should be pointed out that in Harsanyi's original formulation, ethical preferences of an impartial observer are defined over the set $\Delta(I \times X)$. Our modeling strategy of defining it over the set $\Delta(I) \times \Delta(X)$ follows Grant, Kajii, Polak, and Safra (2010).

# 3 A Representation of Subjective Preferences

Our first task is to theorize how individuals whose subjective preferences are sensitive to procedural fairness concerns are accommodated within our framework. To that end, let us divide the set of individuals in set $I$ into ones whose subjective preferences are of the standard vonNeumann and Morgenstern (vNM) type with an expected utility representation and the ones whose are not. Let $I_0 = \{i \in I: \succcurlyeq_i \text{ is vNM}\}$[9] and $I_1 = I \setminus I_0$. Note that, if $\succcurlyeq_i$ such that $\succ_i = \emptyset$, then $\succcurlyeq_i$ is a vNM preference and $i \in I_0$. Therefore, if $i \in I_1$, then $\succ_i \neq \emptyset$. As clarified in the Introduction, individuals in $I_0$ have no concerns for procedural fairness. On the other hand, the key question that we need to answer for the individuals in $I_1$ is the following: *How can we ascertain that procedural fairness concerns is the reason why their subjective preferences depart from Bayesian rationality?* The way we answer this question is by providing an axiomatic foundation in terms of our primitive preference relations that allows us to represent the subjective preferences of these individuals by a Karni-Safra representation (Karni and Safra, 2002). This representation specifically models preferences of individuals who care about procedural fairness in an economic environment that is identical to ours—i.e., one of allocating an indivisible good amongst contesting claimants. In the analysis below, we propose a way of using our primitive preference information to identify and separate out concerns for procedural fairness from concerns for the allocation risk when it comes to these individuals' assessments of outcome lotteries. Specifically, we show using the Karni-Safra representation how these two concerns are distinct dimensions driving preference judgments under their subjective preferences.

Essentially this theorizing involves three key ideas that our axioms will formally clarify. First, when attention is restricted to identity-outcome lotteries for which the identity lottery is a degenerate one with the individual in his position as an impartial observer assuming his own identity for sure, there is a congruence between his ranking of outcome lotteries under his subjective and ethical preferences. As such, attitudes towards risk and procedural fairness embedded in his subjective preferences naturally project on to his ethical preferences over this restricted domain in which no interpersonal comparisons are involved. Second, we maintain that under an individual's ethical preferences, when it comes to assessing the iden-

[9]When we say $\succcurlyeq_i$ is vNM, we, of course, mean that $\succcurlyeq_i$ satisfies the three axioms of (i) completeness and transitivity, (ii) continuity and (iii) independence.

tity risks that he faces under identity-outcome lotteries, he very much behaves like a Bayesian. We use this fact along with the assumption that the source of risk does not influence an individual's attitude towards it to elicit a vNM preference ranking that captures his risk assessments of outcome lotteries. Third, from the information about an individual's overall assessment of an outcome lottery and his risk assessment of it, both measured along his ethical preference scale, we back out as a "residual" another preference ranking that captures his non-Bayesian or procedural assessment of outcome lotteries. We then introduce an axiom that clarifies precisely when we can think of this residual as reflecting concerns for procedural fairness. Once we have teased out, thus, these two concerns for risk and procedural fairness as distinct considerations, we show that we can represent such an individual's subjective preferences as a monotone function of the two, à la Karni and Safra (2002).

We now introduce a set of axioms that lay the groundwork for this exercise. Although our goal in this section is to provide an axiomatic basis to represent the subjective preferences of individuals in $I_1$, the three axioms in the next sub-section applies to all individuals in $I$ as these axioms play a key role when it comes to representing the ethical preferences of individuals via utilitarian or generalized utilitarian SWFs.

## 3.1 Independence, Self-Acceptance and Interpersonal Conflict

Our first axiom introduces a version of the vN-M *independence* condition on an individual's ethical preferences. Specifically, it requires these preferences to adhere to Bayesian rationality when it comes to assessing identity risks from the perspective of as impartial observer. This version of independence that we impose on the ethical preferences of an impartial observer follows Grant, Kajii, Polak, and Safra (2010).

**Axiom 3.1** (Independence Over Identity Lotteries)**.** For $i \in I$, if $(z, l)$, $(z', l') \in \Delta(I) \times \Delta(X)$ are such that $(z, l) \sim_i^* (z', l')$, then for any $\widetilde{z}, \widetilde{z}' \in \Delta(I)$ and $\alpha \in [0, 1]$,

$$(\widetilde{z}, l) \succcurlyeq_i^* (\widetilde{z}', l') \text{ if and only if } (\alpha\widetilde{z} + (1-\alpha)z, l) \succcurlyeq_i^* (\alpha\widetilde{z}' + (1-\alpha)z', l').$$

To understand the content of this axiom, recall our earlier observation that any identity-outcome lottery in $\Delta(I) \times \Delta(X)$ can be equivalently viewed in terms of the product measure in $\Delta(I \times X)$ corresponding to it. Let $(z, l)^*$, $(z', l')^*$, $(\widetilde{z}, l)^*$ and $(\widetilde{z}', l')^*$ denote the product measures in $\Delta(I \times X)$ corresponding to $(z, l)$, $(z', l')$, $(\widetilde{z}, l)$ and $(\widetilde{z}', l')$, respectively. Further, since the outcome lottery is the same under $(z, l)$ and $(\widetilde{z}, l)$, it follows that the lottery $\alpha(\widetilde{z}, l)^* + (1 - \alpha)(z, l)^*$ in $\Delta(I \times X)$ is the product measure corresponding to $(\alpha\widetilde{z} + (1 - \alpha)z, l)$. Similarly, $\alpha(\widetilde{z}', l')^* + (1-\alpha)(z', l')^*$ is the product measure corresponding to $(\alpha\widetilde{z}' + (1-\alpha)z', l')$. Accordingly, for the type of identity-outcome lotteries under consideration, this axiom has the usual interpretation of vN-M independence. That is, in his position as an impartial observer, if $i$ is indifferent between $(z, l)^*$ and $(z', l')^*$, then he should prefer $(\widetilde{z}, l)^*$ to $(\widetilde{z}', l')^*$ if and only if he prefers $\alpha(\widetilde{z}, l)^* + (1 - \alpha)(z, l)^*$ to $\alpha(\widetilde{z}', l')^* + (1 - \alpha)(z', l')^*$; i.e., if he is indifferent between $(z, l)$ and $(z', l')$, then he should prefer $(\widetilde{z}, l)$ to $(\widetilde{z}', l')$ if and only if he prefers $(\alpha\widetilde{z} + (1 - \alpha)z,\ l)$ to $(\alpha\widetilde{z}' + (1 - \alpha)z',\ l')$. In other words, this axiom requires an impartial observer's preferences to satisfy vN-M independence when it comes to facing identity risk.

Our second axiom says that an individual's ethical preferences agree with his subjective preferences when he considers identity-outcome lotteries in which he faces no identity-risk and is guaranteed to be himself with probability one.

**Axiom 3.2** (Self-Acceptance)**.** *For $i \in I$, and any $l, l' \in \Delta(X)$, $l \succcurlyeq_i l'$ if and only if $([i], l) \succcurlyeq_i^* ([i], l')$.*

Our next axiom conveys the idea that an individual's ethical preferences acknowledge the fact that contesting claims on the scarce resource (the indivisible good) make interpersonal conflicts inevitable. It says that if by his ethical preferences, $i$ maintains that individual $j$ is strictly better off under some outcome lottery $l$ than under some other outcome lottery $l'$, then he has to acknowledge that there exists some individual $k$ who is worse off under $l$ than under $l'$. Further, when such interpersonal conflicts exist, it may not be possible for $i$, by his ethical preferences, to give $j$'s claim outright precedence over that of $k$.

**Axiom 3.3** (Interpersonal Conflict)**.** *For $i \in I$, if $l, l' \in \Delta(X)$ are such that $([j], l) \succ_i^* ([j], l')$, for some $j \in I$, then there exists $k \in I$ satisfying (a) $([k], l') \succ_i^* ([k], l)$ and (b) either $([k], l') \succcurlyeq_i^* ([j], l)$ or $([j], l') \succcurlyeq_i^* ([k], l)$.*

Condition (*b*) clarifies why it may not be possible for $i$ to give precedence to $j$

in this situation. This is best understood by observing what is true when this condition does not hold. In this case, $i$'s ethical assessment is given by: $([j], l) \succ_i^* ([k], l') \succ_i^* ([k], l) \succ_i^* ([j], l')$. If this were so then, presumably, $i$ would be justified in maintianing that $j$'s claim should take precedence since the variation in $j$'s well-being, as a result of whether $l$ or $l'$ is chosen, is clearly greater than that of $k$. By ruling out this possibility, condition $(b)$ emphasizes that $i$'s ethical preferences preclude giving outright precedence to anyone's claim when such interpersonal conflicts exist.

## 3.2 Separating Risk and Procedural Fairness Concerns

Consider any individual $i \in I_1$. For any such individual, our goal now is to propose a way of identifying and separating out his concerns for procedural fairness from his risk concerns when it comes to assessing outcome lotteries. The starting point of this exercise is to draw on the self-acceptance axiom. Because of this axiom, when attention is restricted to identity-outcome lotteries under which the identity lottery guarantees that, in his position as an impartial observer, the individual is himself with probability one, the ranking of outcome lotteries implied by his ethical preferences is identical to that under his subjective preferences. Therefore, any attitudes towards risk and procedural fairness that are embedded in his subjective preferences naturally project on to his ethical preference scale on this restricted domain, $\{i\} \times \Delta(X)$, in which no inter-personal comparisons are involved. In other words, when it comes to doing this separation of risk and procedural fairness concerns there is a congruence between doing the exercise from the perspective of his ethical and subjective preferences. We will start by doing this exercise from the perspective of the former. That is, for any individual $i \in I_1$, we will focus on his ethical preferences $\succcurlyeq_i^*$ and back out his risk and procedural fairness attitudes over outcome lotteries. We will then show that these attitudes very much drive behavior under his subjective preferences—a statement that we will formalize by means of representing subjective preferences through a Karni-Safra representation under which subjective preferences are monotone in these attitudes.

We first focus on identifying risk attitudes. With that goal, consider the following definition.

**Definition 3.1.** *Let $i \in I_1$, $l = [x_1, \alpha_1; \ \ldots \ ; x_M, \alpha_M] \in \Delta(X)$ be such that there*

*exists* $l' \in \Delta(X)$ *and* $z_m \in \Delta(I)$ *satisfying* $([i],[x_m]) \sim_i^* (z_m, l')$ *for each* $m = 1, \ldots, M$. *Then, we call* $(\alpha_1 z_1 + \ \ldots \ + \alpha_M z_M, l')$ *a* ***risk equivalent*** *of* $l$ *for* $i$.

Consider the outcome lottery $l = [x_1, \alpha_1; \ldots; x_M, \alpha_M] \in \Delta(X)$. What is individual $i$'s assessment of the allocation risk under it when measured along his ethical preference scale? The above definition provides an answer to this question by proposing a candidate identity-outcome lottery on this scale that may be identified with his assessment of the allocation risk under $l$. Specifically, since for each $x_m$ in the support of $l$, $([i],[x_m]) \sim_i^* (z_m, l')$, his assessment of the allocations $x_1, \ldots, x_M$ can, respectively, be identified with his assessment of the identity-outcome lotteries $(z_1, l'), \ldots, (z_M, l')$ on his ethical preference scale. As such, given that his ethical preference relation satisfies the independence over identity lotteries axiom, his assessment of the allocation risk under $l$ can be identified with his assessment, according to $\succcurlyeq_i^*$, of the second-order identity risk under the identity-outcome lottery $(\alpha_1 z_1 + \ \ldots \ + \alpha_M z_M, l')$. That is, assuming that the source of the risk doesn't influence his attitude towards similar risks, we may think of $(\alpha_1 z_1 + \ \ldots \ + \alpha_M z_M, l')$ as a risk equivalent of the outcome lottery $l$ for $i$. Observe that for any degenerate outcome lottery $[x]$, any $(z, l)$ s.t. $(z, l) \sim_i^* ([i],[x])$ is a risk equivalent of $[x]$ for $i$, including $([i],[x])$.

We can use the notion of a risk equivalent to define, for each $i \in I_1$, a preference (binary) relation $\succcurlyeq_i^R \subseteq \Delta(X) \times \Delta(X)$ that captures $i$'s risk attitudes over outcome lotteries. Specifically, for any $l, l' \in \Delta(X)$, $l \succcurlyeq_i^R l'$ if $(\tilde{z}, \tilde{l}) \succcurlyeq_i^* (\tilde{z}', \tilde{l}')$, where $(\tilde{z}, \tilde{l})$ and $(\tilde{z}', \tilde{l}')$ are, respectively, risk equivalents of $l$ and $l'$ for $i$. It is straightforward to verify that $\succcurlyeq_i^R$ is well-defined: if $(\tilde{z}, \tilde{l})$ and $(\hat{z}, \hat{l})$ are both risk equivalents of $l$, then by virtue of the independence over identity lotteries axiom, $(\tilde{z}, \tilde{l}) \sim_i^* (\hat{z}, \hat{l})$. Under our axioms, the following result follows:

**Proposition 3.1.** *If* $i \in I_1$ *satisfies independence over identity lotteries and interpersonal conflict, then* $\succcurlyeq_i^R$ *is a vNM preference relation.*[10]

Now that we have ascertained, using the notion of a risk equivalent, what $i$'s assessment of the allocation risk under an outcome lottery is, we proceed to identify a non-Bayesian or procedural preference relation for him that captures his assessments of outcome lotteries when viewed in their role as allocation procedures. The way we do this is the following. Consider two outcome lotteries, $l$ and $l'$, and for

[10] That is, $\succcurlyeq_i^R$ is complete, transitive, and satisfies continuity and independence.

the sake of the exposition, suppose that the overall assessment of the two identity-outcome lotteries $([i], l)$ and $([i], l')$ on his ethical preference scale correspond to 100 and 200 "utils," respectively. Further, suppose that his assessment of the allocation risk under these lotteries as measured by their risk equivalents on this same scale correspond to 70 and 130 utils, respectively. In other words, his assessment of the difference between these two outcome lotteries $(200 - 100 = 100)$ cannot be explained based solely on the difference in his assessment of the allocation risk under them $(130 - 70 = 60)$. There is a positive residual $(100 - 60 = 40 > 0)$ that needs to be accounted for. This residual reveals the fact that distinct from the difference between these two outcome lotteries based on their risk assessments, there is an additional source of difference that has to do, presumably, with their role as allocation procedures and on this dimension $l$ has an advantage over $l'$. This being the case, $l$ should rank higher than $l'$ under this procedural preference relation. How do we formalize this argument? It turns out that under our axioms there is a way of doing so purely in terms of preference information. Indeed, the axioms of independence over identity lotteries and interpersonal conflict imply that there is a representation of ethical preferences that is cardinal in the vNM sense.[11] Hence, notions of utility difference comparisons as articulated above have meaning in our setting and can be formalized in terms of primitive preferences.

**Definition 3.2.** *For $i \in I_1$, the procedural preference (binary) relation $\succcurlyeq_i^P \subseteq \Delta(X) \times \Delta(X)$ is defined as: for any $l, l' \in \Delta(X)$, $l \succcurlyeq_i^P l'$ if there exists $l^* \in \Delta(X)$ and $z$, $z'$, $\widetilde{z}$, $\widetilde{z}' \in \Delta(I)$ such that:*

1. *$(z, l^*)$ and $(z', l^*)$ are, respectively, risk equivalents of $l$ and $l'$ for $i$;*
2. *$([i], l) \sim_i^* (\widetilde{z}, l^*)$ and $([i], l') \sim_i^* (\widetilde{z}', l^*)$; and*
3. *$(.5\widetilde{z} + .5z', l^*) \succcurlyeq_i^* (.5\widetilde{z}' + .5z, l^*)$.*

*We denote the symmetric and asymmetric components of $\succcurlyeq_i^P$ by $\sim_i^P$ and $\succ_i^P$, respectively.*

The reasoning behind the definition is the following. The preferences $([i], l) \sim_i^* (\widetilde{z}, l^*)$ and $([i], l') \sim_i^* (\widetilde{z}', l^*)$ imply that the identity-outcome lotteries $(\widetilde{z}, l^*)$ and $(\widetilde{z}', l^*)$ are, respectively, the projections on $i$'s ethical preference scale of his overall

[11] This is established in Lemma A.2 in the Appendix.

assessments of the outcome lotteries $l$ and $l'$. Additionally, $(z, l^*)$ and $(z', l^*)$ are, respectively, the risk equivalents of $l$ and $l'$ and capture the risk assessments of these outcome lotteries on this scale. Therefore, drawing on the independence over identity lotteries axiom, it follows that the identity-outcome lottery $(.5\widetilde{z} + .5z', l^*)$ is akin to a 50:50 mixture of the overall assessment of $l$ and the risk assessment of $l'$. Similarly, $(.5\widetilde{z'} + .5z, l^*)$ is akin to a 50:50 mixture of the overall assessment of $l'$ and the risk assessment of $l$. If all that entered $i$'s overall assessments of $l$ and $l'$ was a consideration for their risk assessments, then he ought to be indifferent between $(.5\widetilde{z} + .5z', l^*)$ and $(.5\widetilde{z'} + .5z, l^*)$ as, in that case, both would simply be equivalent to a 50:50 mixture of his risk assessments of $l$ and $l'$. On the other hand, if he expresses a preference for $(.5\widetilde{z} + .5z', l^*)$ over $(.5\widetilde{z'} + .5z, l^*)$, then it reveals the fact that, beyond their risk assessments, $i$ considers $l$ to have an advantage over $l'$ on the ground that it is a preferable allocation procedure. Finally, note that by virtue of independence over identity lotteries, the binary relation $\succcurlyeq_i^P$ is well-defined.

The following proposition establishes that under our axioms $\succcurlyeq_i^P$ is a weak order.

**Proposition 3.2.** *If $i \in I_1$ satisfies independence over identity lotteries and interpersonal conflict, then $\succcurlyeq_i^P$ is complete and transitive.*

The procedural preference relation captures an individual's concerns for procedural fairness in evaluating outcome lotteries, if such attitudes are present in his ethical/subjective preferences. However, in principle, it could also capture other deviations from Bayesian rationality in these preferences. How can we maintain that any deviation from Bayesian rationality that it identifies is attributable to concerns for procedural fairness alone? The following axiom, by identifying a structure on this preference relation, helps us address this question.

**Axiom 3.4** (Revealed Fairness)**.** *For $i \in I_1$, if $l, l' \in \Delta(X)$ are such that $l \sim_i^P l'$, then for any $\alpha \in (0, 1)$, $\alpha l + (1 - \alpha) l' \succ_i^P l$.*

The axiom connects the procedural preference relation to the key rationale as to why it may be desirable, on grounds of procedural fairness, to allocate an indivisible good by means of a lottery. Specifically, it draws on the well-known insight that randomizing between outcome lotteries can play a crucial role in enhancing procedural fairness in the context of using such lotteries to allocate an indivisible good. Presumably, such a judgment is based on the observation

that if we consider two lotteries, $l$ and $l'$, then $l$ may provide more favorable opportunities than $l'$ for some individuals, whereas the opposite may be true for others. Accordingly, randomizing between these lotteries may provide a way to balance these contesting claims and arrive at a fairer allocation procedure. The axiom requires the procedural preference relation to inherit this reasoning. Hence, it states that when an individual finds two outcome lotteries to be equally good allocation procedures, he must find their mixture to be a strictly better allocation procedure. This axiom and its justification is identical to the *compromise fairness* axiom of Karni and Safra (2002).

The exercise of deriving the risk and procedural preference relations above was done based on assessments made along the ethical preference scale of an individual. The self-acceptance axiom and the discussion at the beginning of this sub-section suggests that these preference relations should also have a close connection to the individual's subjective preferences. Intuitively speaking, this axiom implies that attitudes that are present in an individual's subjective preferences ought to find faithful expression in his ethical preferences when attention is restricted to those situations where, as an impartial observer, he does not face any interpersonal conflict and can fully subscribe to his own subjective preferences. Therefore, it stands to reason that the risk and procedural preference relations that we derived does indeed capture deep features of his subjective preferences. We now formally establish this observation. The Proposition below establishes that the risk and procedural preference relations are indeed distinct dimensions that drive preference judgments under his subjective preferences.

**Proposition 3.3.** *If $i \in I_1$ satisfies independence over identity lotteries, self-acceptance and interpersonal conflict, then for all $l, l' \in \Delta(X)$,*

1. $l \sim_i^R l' \Longrightarrow [l \succcurlyeq_i l' \text{ iff } l \succcurlyeq_i^P l']$
2. $l \sim_i^P l' \Longrightarrow [l \succcurlyeq_i l' \text{ iff } l \succcurlyeq_i^R l']$

*In addition, if $i$ satisfies revealed fairness, then for all $l, l' \in \Delta(X)$, $l \neq l'$,*

$$[l \sim_i l' \text{ and } l \sim_i^R l'] \Longrightarrow \alpha l + (1-\alpha) l' \succ_i l$$

The final part of the Proposition clarifies precisely when we will see a strict preference for randomization in subjective preferences. It tells us that the logic for such

randomization emanates precisely from procedural preference concerns and not risk concerns. Note that since $\succcurlyeq_i^R$ is vNM, $l \sim_i^R l'$ implies that $\alpha l + (1-\alpha) l' \sim_i^R l$. We now further extend the message of this Proposition and show that subjective preferences have a Karni-Safra representation. In our setting, this representation expresses assessments of outcome lotteries as an aggregation of the two distinct considerations of risk and procedural fairness captured by the preference relations $\succcurlyeq_i^R$ and $\succcurlyeq_i^P$, respectively.

## 3.3 Karni-Safra Representation

In the way of notation, note that for any function $u_i : X \to \mathbb{R}$, we will denote the expected utility functional with respect to it by $\mathbb{E}u_i : \Delta(X) \to \mathbb{R}$, given by $\mathbb{E}u_i(l) = \sum_{x \in X} l(x) u_i(x)$.

**Definition 3.3.** *A Karni-Safra (KS) representation of $\succcurlyeq_i$, $i \in I_1$, consists of three functions*

1. $u_i : X \to \mathbb{R}$*;*
2. $g_i : \Delta(X) \to \mathbb{R}$ *that is continuous and strictly quasiconcave with* $g_i([x]) = 0$ *for any* $x \in X$*; and*
3. $\psi_i : \{(\mathbb{E}u_i(l), g_i(l)) : l \in \Delta(X)\} \to \mathbb{R}$ *that is strictly increasing;*

*such that*

1. $u_i$ *is a vNM representation of* $\succcurlyeq_i^R$*, i.e., for any* $l, l' \in \Delta(X)$*,* $l \succcurlyeq_i^R l'$ *iff* $\mathbb{E}u_i(l) \geq \mathbb{E}u_i(l')$*;*
2. $g_i$ *represents* $\succcurlyeq_i^P$*; and*
3. *the function* $U_i : \Delta(X) \to \mathbb{R}$*, given by* $U_i(l) = \psi_i(\mathbb{E}u_i(l), g_i(l))$ *represents* $\succcurlyeq_i$*.*

Under a KS representation of $\succcurlyeq_i$, $i \in I_1$, there exists a function $u_i$ that captures $i$'s risk attitudes embodied in the preference relation $\succcurlyeq_i^R$. Specifically, $\succcurlyeq_i^R$ has an

expected utility representation with $u_i$ as the Bernoulli utility function. Further, there exists a function $g_i$ that captures $i$'s concern for the allocation procedure. For any $l \in \Delta(X)$, $g_i(l)$ captures $i$'s assessment of how fair this outcome lottery is as an allocation procedure. The strict quasi-concavity of the $g_i$ function provides the room to accommodate a preference for randomization owing to concerns for procedural fairness. Finally, the overall utility assessment of any outcome lottery $l$ can be expressed as an increasing function of the expected utility component, $\mathbb{E}u_i(l)$, and the procedural fairness component, $g_i(l)$.

A special kind of KS representation that plays an important role in our analysis is what we refer to as a *basic KS representation*. A KS representation $(u_i, g_i, \psi_i)$ is basic if for any $l \in \Delta(X)$,

$$\psi_i(\mathbb{E}u_i(l), g_i(l)) = \mathbb{E}u_i(l) + g_i(l)$$

We denote a basic KS representation as a pair $(u_i, g_i)$.

The following result establishes that under our axioms, for any $i \in I_1$, *every* representation of $\succcurlyeq_i$ is a KS representation. Of course, given that $\succcurlyeq_i$ is a continuous weak order, it has a utility representation. In addition, the axioms also guarantee that $\succcurlyeq_i$ has a basic KS representation.

**Theorem 3.1.** *If $i \in I_1$ satisfies independence over identity lotteries, self-acceptance, interpersonal conflict and revealed fairness then (i) every representation of $\succcurlyeq_i$ is a KS representation, and (ii) $\succcurlyeq_i$ has a basic KS representation.*

# 4 Representation of Ethical Preferences

We are now ready to accomplish the primary task of this paper, which is to provide a foundation for the statement that procedural fairness concerns can be elicited from individual subjective preferences and incorporated into social welfare judgments. As we have seen, for individuals in the set $I_1$ who care about procedural fairness, a KS representation of subjective preferences allows us to identify and represent such concerns. On the other hand, for individuals in the set $I_0$ who do not care about procedural preferences, subjective preferences have a standard

expected utility representation. Our goal here is to show how all such individual assessments can be faithfully incorporated within any social welfare function (SWF) representing the ethical preferences of an impartial observer. We show that this can be done for both utilitarian SWFs as well as generalized utilitarian ones. We begin with the latter class of SWFs.

## 4.1 Generalized Utilitarianism

We first formally define our notion of a generalized utilitarian SWF that incorporates individual attitudes towards procedural fairness. As mentioned above, our notion of any SWF is a subjectivist one and we view any such function as a representation of the ethical preferences of some individual in society. That is, it captures such an individual's welfare judgments from his perspective as an impartial observer.

**Definition 4.1.** *The collection of ethical preference relations $(\succcurlyeq_i^*)_{i\in I}$ admit generalized utilitarian representations that incorporate individuals' sense of justice if there exists a collection of functions, $(u_i)_{i\in I_0}$, $(u_i, g_i, \psi_i)_{i\in I_1}$, $((\phi_{ij} : \mathbb{R} \to \mathbb{R})_{j\in I})_{i\in I}$, such that for each $i \in I$,*

1. *if $i \in I_0$, then $u_i$ is a vNM representation of $\succcurlyeq_i$; and if $i \in I_1$, then $(u_i, g_i, \psi_i)$ is a KS representation of $\succcurlyeq_i$. That is, the function $U_i : \Delta(X) \to \mathbb{R}$, given by*

$$U_i(l) = \begin{cases} \mathbb{E}u_i(l), & \text{if } i \in I_0 \\ \psi_i(\mathbb{E}u_i(l), g_i(l)), & \text{if } i \in I_1 \end{cases}$$

   *represents $\succcurlyeq_i$;*

2. *$\phi_{ij}$ is an increasing function for each $j \in I$; and*

3. *the function $V_i : \Delta(I) \times \Delta(X) \to \mathbb{R}$, given by*

$$V_i(z, l) = \textstyle\sum_{j\in I} z(j)\phi_{ij}(U_j(l)),$$

   *represents $\succcurlyeq_i^*$.*

Observe that under a generalized utilitarian representation, each individual's assessment of any outcome lottery is based on either a KS representation or an expected utility representation depending on whether the individual cares about procedural fairness or not. Further, when representing the ethical preferences of individual $i$, the function $\phi_{ij}$ translates individual $j$'s utility scale into individual $i$'s. It is in this sense that this representation generalizes a utilitarian representation under which no such translation is admissible (as we will see below).

We need to introduce one additional axiom, which in conjunction to the ones above, provides a foundation for a generalized utilitarian representation. This axiom is Harsanyi's acceptance principle. In our set-up, it says that if individual $i$, from his perspective as an impartial observer, knows for sure that he will assume individual $j$'s identity, then his ethical preferences should coincide with that of $j$'s subjective preferences. Observe that the acceptance principle implies that each individual's preferences satisfy self-acceptance.

**Axiom 4.1** (Acceptance Principle)**.** *For $i \in I$, and any $l, l' \in \Delta(X)$, $j \in I$, $l \succcurlyeq_j l'$ if and only if $([j], l) \succcurlyeq_i^* ([j], l')$.*

**Theorem 4.1.** *Suppose each $i \in I$ satisfies interpersonal conflict. Then:*

1. *The collection of ethical preferences $(\succcurlyeq_i^*)_{i\in I}$ admit generalized utilitarian representations that incorporate individuals' sense of justice if and only if each $i \in I$ satisfies independence over identity lotteries and the acceptance principle and, in addition, each $i \in I_1$ satisfies revealed fairness.*

2. *If $((u_i)_{i\in I_0}, (u_i, g_i, \psi_i)_{i\in I_1}, ((\phi_{ij})_{j\in I})_{i\in I})$ and $((\widetilde{u}_i)_{i\in I_0}, (\widetilde{u}_i, \widetilde{g}_i, \widetilde{\psi}_i)_{i\in I_1}, ((\widetilde{\phi}_{ij})_{j\in I})_{i\in I})$ are both generalized utilitarian representations of $(\succcurlyeq_i^*)_{i\in I}$ that incorporate individuals' sense of justice then, for each $i \in I$, there exist constants $\tau_i > 0$, $\tau_i'$ such that $\widetilde{\phi}_{ij} \circ \widetilde{U}_j = \tau_i(\phi_{ij} \circ U_j) + \tau_i'$, for all $j \in I$ where $U_j, \widetilde{U}_j : \Delta(X) \to \mathbb{R}$ are given by*

$$U_j(l) = \begin{cases} \mathbb{E}u_j(l), & \text{if } j \in I_0 \\ \psi_j(\mathbb{E}u_j(l), g_j(l)), & \text{if } j \in I_1 \end{cases}$$

$$\widetilde{U}_j(l) = \begin{cases} \mathbb{E}\widetilde{u}_j(l), & \text{if } j \in I_0 \\ \psi_j(\mathbb{E}\widetilde{u}_j(l), \widetilde{g}_j(l)), & \text{if } j \in I_1 \end{cases}$$

## 4.2 Utilitarianism

We next define what it means for individuals' ethical preferences, reflecting welfare judgments, to have utilitarian SWF representations. In such a representation, when it comes to representing the subjective preferences of individuals who care about procedural fairness, we restrict attention to basic KS representations.

**Definition 4.2.** *The collection of ethical preference relations $(\succcurlyeq_i^*)_{i\in I}$ admit utilitarian representations that incorporate individuals' sense of justice if there exists a collection of functions, $(u_i)_{i\in I_0}$, $(u_i, g_i)_{i\in I_1}$, such that for each $i \in I$,*

1. *if $i \in I_0$, then $u_i$ is a vNM representation of $\succcurlyeq_i$; and if $i \in I_1$, then $(u_i, g_i)$ is a basic KS representation of $\succcurlyeq_i$. That is, the function $U_i : \Delta(X) \to \mathbb{R}$, given by*

$$U_i(l) = \begin{cases} \mathbb{E}u_i(l), & \text{if } i \in I_0 \\ \mathbb{E}u_i(l) + g_i(l), & \text{if } i \in I_1 \end{cases}$$

   *represents $\succcurlyeq_i$; and*

2. *the function $V : \Delta(I) \times \Delta(X) \to \mathbb{R}$ given by*

$$V(z, l) = \textstyle\sum_{j\in I} z(j)U_j(l)$$

   *represents $\succcurlyeq_i^*$.*

Observe one stark property of welfare judgments under such a representation. There must necessarily be unanimity in society about such judgments. Therefore, the following axiom—which says that individuals in society agree on their preferences over identity-outcome lotteries when they view things impersonally—is necessary for a utilitarian representation.

**Axiom 4.2** (Shared Ethics)**.** *For all $i, j \in I$, $(z, l), (z', l') \in \Delta(I)\times\Delta(X)$, $(z, l) \succcurlyeq_i^* (z', l')$ if and only if $(z, l) \succcurlyeq_j^* (z', l')$.*

In addition, a utilitarian representation imposes a certain kind of consistency on attitudes towards randomization of any individual whose subjective preferences are of the vNM type, i.e., any individual $i \in I_0$. This consistency requirement

is that any such individual, when faced with randomization in his environment, should not distinguish between the source of the randomization and assess similar randomizations similarly. Specifically, under his ethical preferences, when presented with a randomization over outcome lotteries and a similar randomization over identity lotteries, he must be indifferent between the two. The reasoning behind this is the following. Given that procedural fairness concerns do not enter such an individual's subjective assessments, any randomization over outcome lotteries influences such assessments due to risk considerations alone. Further, if the welfare criterion is a utilitarian one, his subjective utility scale incorporating these risk assessments must be inherited one-to-one in his ethical assessments as an impartial observer. In turn, this means that his ethical assessments can no longer discriminate between similar randomizations over outcome and identity lotteries and he is forced to be indifferent between the two. This idea that the source of randomization should not influence attitudes towards it is at the core of Harsanyi's utilitarianism and its formalization in the current context is the following.[12]

**Axiom 4.3** (Indifference to Similar Randomizations)**.** *For $i \in I_0$, $l, l', \tilde{l} \in \Delta(X)$ and $z, z' \in \Delta(I)$, if $([i], l) \sim_i^* (z, \tilde{l})$ and $([i], l') \sim_i^* (z', \tilde{l})$, then for all $\alpha \in [0,1]$, $([i], \alpha l + (1-\alpha)l') \sim_i^* (\alpha z + (1-\alpha)z', \tilde{l})$.*

The following result establishes that these two axioms, along with ones specified earlier, are sufficient for a utilitarian representation.

**Theorem 4.2.** *Suppose each $i \in I$ satisfies interpersonal conflict. Then:*

1. *The collection $(\succcurlyeq_i^*)_{i \in I}$ admit utilitarian representations that incorporate individuals' sense of justice if and only if each $i \in I$ satisfies independence over identity lotteries and self-acceptance, each $i, j \in I$ satisfies shared ethics, and, in addition, each $i \in I_1$ satisfies revealed fairness and each $i \in I_0$ satisfies indifference to similar randomizations.*

2. *If $((u_i)_{i \in I_0}, (u_i, g_i)_{i \in I_1})$ and $((\widetilde{u}_i)_{i \in I_0}, (\widetilde{u}_i, \widetilde{g}_i)_{i \in I_1})$ are both utilitarian representations of $(\succcurlyeq_i^*)_{i \in I}$ that incorporate individuals' sense of justice, then there exist constants $\tau > 0$, $\tau'$ such that $\widetilde{u}_i = \tau u_i + \tau'$ for all $i \in I$ and $\widetilde{g}_i = \tau g_i$, for all $i \in I_1$.*

[12]The spirit of this axiom is similar to the Indifference Between Life Chances and Accidents of Birth axiom in Grant, Kajii, Polak, and Safra (2010).

Before concluding, a couple of comments are in order.[13]

1. Under a utilitarian SWF representation of ethical preferences, there is no scope for an impartial observer to independently add an intensity for procedural fairness in welfare assessments beyond what individual attitudes towards procedural fairness—elicited from their subjective preferences—demand. In particular, if all of the individual subjective preferences reveal an indifference towards procedural fairness concerns, then social welfare assessments cannot express a strict preference for procedural fairness. However, this need not be the case under a generalized utilitarian SWF. Under generalized utilitarianism, an impartial observer's preferences may be an additional source of procedural fairness concerns in welfare assessments over and above what is dictated by individual subjective preferences. Specifically, it may be possible for social welfare judgments to exhibit a concern for procedural fairness even when all the individual subjective preferences show no concern for it. Viewing individual $i$ as the impartial observer, this may be the case when all the functions $\phi_{ij}$ are strictly concave.

2. Another feature of a utilitarian representation of ethical preferences worth highlighting is that, under it, all individuals in society have to exhibit identical attitudes towards risk. In other words, under utilitarianism, it is not possible to accommodate the feature that one individual might be more comfortable facing a risk than another individual. This is not the case under generalized utilitarianism where different individual risk attitudes can be accommodated.

# A Proofs

## A.1 Preliminary Results

We begin with some preliminary results. To that end, for any $i \in I$ and $l \in \Delta(X)$, define $\succcurlyeq_{i,l} \subseteq \Delta(I) \times \Delta(I)$ as follows: $z \succcurlyeq_{i,l} z'$ if $(z, l) \succcurlyeq_i^* (z', l)$. Let $\succ_{i,l}$ and $\sim_{i,l}$

[13]For a more detailed discussion of these points, please refer to Grant, Kajii, Polak, and Safra (2010).

denote the asymmetric and symmetric components of $\succcurlyeq_{i,l}$, respectively. Further, observe that since $\succcurlyeq_i^*$ is a continuous weak order and satisfies independence over identity lotteries, $\succcurlyeq_{i,l}$ satisfies the three vNM axioms, including vNM independence.

**Lemma A.1.** *Suppose $i \in I$ satisfies independence over identity lotteries and interpersonal conflict and let $l' \in \Delta(X)$ be such that $\succ_{i,l'} \neq \emptyset$. Then for any $l \in \Delta(X)$:*

1. *If $\succ_{i,l} = \emptyset$, then there exists $\widetilde{z} \in \Delta(I)$ such that $(\widetilde{z}, l') \sim_i^* (z, l)$, for all $z \in \Delta(I)$.*
2. *If $\succ_{i,l} \neq \emptyset$, then there exists $\widetilde{z}$, $\widehat{z}$, $\widetilde{z}'$, $\widehat{z}' \in \Delta(I)$ such that $(\widetilde{z}, l) \sim_i^* (\widetilde{z}', l') \succ_i^* (\widehat{z}, l) \sim_i^* (\widehat{z}', l')$.*

*Proof.* Consider $l \in \Delta(X)$ for which $\succ_{i,l} = \emptyset$, i.e., $(z, l) \sim_i^* (z', l)$ for all $z$, $z' \in \Delta(I)$. To establish our desired conclusion for this case, note that, since $\succ_{i,l'} \neq \emptyset$ and $\succcurlyeq_i^*$ satisfies independence over identity lotteries, there exists $j', j'' \in I$ such that $([j'], l') \succ_i^* ([j''], l')$. Further, since $([j'], l) \sim_i^* ([j''], l)$, it follows that there exists $j = j'$ or $j''$ (possibly both), such that $\neg[([j], l) \sim_i^* ([j], l')]$. Suppose that $([j], l') \succ_i^* ([j], l)$ (the other case of $([j], l) \succ_i^* ([j], l')$ can he handled along similar lines). Then, interpersonal conflict implies that there exists $k \in I$ such that $([j], l) \succcurlyeq_i^* ([k], l')$, since it cannot be the case that $([k], l) \succcurlyeq_i^* ([j], l')$, for this would violate $\succ_{i,l} = \emptyset$. If $([k], l') \sim_i^* ([j], l) \sim_i^* (z, l)$, for all $z \in \Delta(I)$, then we have our desired conclusion. On the other hand, if $([j], l') \succ_i^* ([j], l) \succ_i^* ([k], l')$, it follows from the continuity of $\succcurlyeq_i^*$ and independence over identity lotteries, that there exists $\widetilde{z} \in \Delta(I)$ such that $(\widetilde{z}, l') \sim_i^* ([j], l) \sim_i^* (z, l)$ for all $z \in \Delta(I)$.

Next, consider $l \in \Delta(X)$ for which $\succ_{i,l} \neq \emptyset$, i.e., $(z, l) \succ_i^* (z', l)$ for some $z$, $z' \in \Delta(I)$. To establish our desired conclusion for this case, let $\overline{i}(l)$, $\underline{i}(l) \in I$ be such that $([\overline{i}(l)], l) \succcurlyeq_i^* ([i'], l) \succcurlyeq_i^* ([\underline{i}(l)], l)$ for all $i' \in I$. Clearly, $([\overline{i}(l)], l) \succ_i^* ([\underline{i}(l)], l)$, since $\succcurlyeq_{i,l}$ such that $\succ_{i,l} \neq \emptyset$ and satisfies vN-M independence. Next, note that it cannot be that $([i'], l') \succcurlyeq_i^* ([\overline{i}(l)], l)$ for all $i' \in I$. To see this, suppose this were true. In that case, since $\succ_{i,l'} \neq \emptyset$, there exists $j \in I$ such that $([j], l') \succ_i^* ([\overline{i}(l)], l) \succcurlyeq_i^* ([j], l)$. But, then interpersonal conflict implies that there exists $k \in I$, such that $([k], l) \succ_i^* ([k], l')$. This, in turn, implies that $([k], l) \succ_i^* ([\overline{i}(l)], l)$, contradicting the definition of $\overline{i}(l)$. A similar argument establishes that it cannot be the case that

$([\underline{i}(l)], l) \succcurlyeq_i^* ([i'], l')$ for all $i' \in I$. That is, there exists $j', j'' \in I$, not necessarily distinct, such that $([\overline{i}(l)], l) \succ_i^* ([j'], l')$ and $([j''], l') \succ_i^* ([\underline{i}(l)], l)$. Accordingly, since $\succcurlyeq_i^*$ is continuous and satisfies independence over identity lotteries, we can find $\widetilde{z}$, $\widehat{z}$, $\widetilde{z}'$, $\widehat{z}' \in \Delta(I)$ satisfying $(\widetilde{z}, l) \sim_i^* (\widetilde{z}', l') \succ_i^* (\widehat{z}, l) \sim_i^* (\widehat{z}', l')$. □

The following lemma draws on Grant, Kajii, Polak, and Safra (2010) and Karni and Safra (2000).

**Lemma A.2.** *For any $i \in I$, if $\succ_{i,l'} \neq \emptyset$, for some $l' \in \Delta(X)$, and $i$ satisfies independence over identity lotteries and interpersonal conflict, then for each $l \in \Delta(X)$, there exists a function $v_{i,l} : I \to \mathbb{R}$ such that the function $V_i : \Delta(I) \times \Delta(X) \to \mathbb{R}$ given by $V_i(z, l) = \sum_{j \in I} z(j) v_{i,l}(j)$ represents $\succcurlyeq_i^*$, with the range of $V_i$ a connected set.*[14] *Further, the family of functions $(v_{i,l})_{l \in \Delta(X)}$ is unique up to a common positive affine transformation. That is, if $(\tilde{v}_{i,l})_{l \in \Delta(X)}$ is another family of functions that represents $\succcurlyeq_i^*$ in the above sense, then there exists constants $\tau_i > 0$ and $\tau_i'$ such that $\tilde{v}_{i,l} = \tau_i v_{i,l} + \tau_i'$, for all $l \in \Delta(X)$.*

*Proof.* For any $l \in \Delta(X)$, since $\succcurlyeq_{i,l}$ satisfies the three vNM axioms, there exists a function $v_{i,l} : I \to \mathbb{R}$ such that the function $V_{i,l} : \Delta(I) \to \mathbb{R}$, given by $V_{i,l}(z) = \sum_{j \in I} z(j) v_{i,l}(j)$, represents $\succcurlyeq_{i,l}$. Further, the function $v_{i,l}$ is unique up to a positive affine transformation. We will now piece together the family of $V_{i,l}$ functions to define a function $V_i : \Delta(I) \times \Delta(X) \to \mathbb{R}$ that satisfies the requirements of the lemma. Consider $l' \in \Delta(X)$ for which $\succ_{i,l'} \neq \emptyset$ and begin by defining the function $V_i$ on the set $\Delta(I) \times \{l'\}$ by setting $V_i(z, l') = V_{i,l'}(z)$, for all $(z, l') \in \Delta(I) \times \{l'\}$. Next, we define the function $V_i$ on the sets $\Delta(I) \times \{l\}$ for $l \neq l'$. To do so, we consider the two cases discussed in Lemma A.1.

First, consider those $l \in \Delta(X)$ for which $\succ_{i,l} = \emptyset$. For this case, we know from Lemma A.1 that there exists $\widetilde{z} \in \Delta(I)$ such that $(\widetilde{z}, l') \sim_i^* (z, l)$, for all $z \in \Delta(I)$. Re-define the constant function $V_{i,l}$ by setting $V_{i,l}(z) = V_i(\widetilde{z}, l')$ for all $z \in \Delta(I)$, so that, in particular, $v_{i,l}(j) = V_{i,l}([j]) = V_i(\widetilde{z}, l')$ for all $j \in I$. We can, then, extend the function $V_i$ to $\Delta(I) \times \{l\}$ by defining $V_i(z, l) = V_{i,l}(z)$ for all $(z, l) \in \Delta(I) \times \{l\}$.

Second, consider those $l \in \Delta(X)$ for which $\succ_{i,l} \neq \emptyset$. In this case, we have established in Lemma A.1 that there exists $\widetilde{z}$, $\widehat{z}$, $\widetilde{z}'$, $\widehat{z}' \in \Delta(I)$ such that $(\widetilde{z}, l) \sim_i^*$

[14] Note that such a function $V_i$ is linear in "identity-probabilities." That is, for any $z_1, \ldots, z_M \in \Delta(I)$ and $l \in \Delta(X)$, we have $V_i(\alpha_1 z_1 + \cdots + \alpha_M z_M, l) = \sum_{m=1}^M \alpha_m V_i(z_m, l)$.

$(\widetilde{z}', l') \succ_i^* (\widehat{z}, l) \sim_i^* (\widehat{z}', l')$. Further, recall that the function $V_{i,l}$ is defined uniquely up to a positive affine transformation; that is, we have two degrees of freedom in specifying it. Accordingly, we can *redefine* it by setting $V_{i,l}(\widetilde{z}) = V_i(\widetilde{z}', l')$ and $V_{i,l}(\widehat{z}) = V_i(\widehat{z}', l')$, so that, in particular, we redefine the function $v_{i,l} : I \to \mathbb{R}$ by setting $v_{i,l}(j)$ equal to the 'new' value of $V_{i,l}([j])$ for all $j \in I$. We can, then, extend the function $V_i$ to the set of lotteries in $\Delta(I) \times \{l\}$ by defining $V_i(z, l) = V_{i,l}(z)$ for all $(z, l) \in \Delta(I) \times \{l\}$. This gives us the function $V_i : \Delta(I) \times \Delta(X) \to \mathbb{R}$ as desired in the statement of the lemma. It is fairly straightforward to verify that the function $V_i$ represents the preference relation $\succcurlyeq_i^*$. Further, note that the range of this function is a connected set.

To prove the second part of the lemma, let $(v_{i,l})_{l \in \Delta(X)}$ and $(\tilde{v}_{i,l})_{l \in \Delta(X)}$ be two such representations of $\succcurlyeq_i^*$. Define the functions $V_i : \Delta(I) \times \Delta(X) \to \mathbb{R}$ and $\widetilde{V}_i : \Delta(I) \times \Delta(X) \to \mathbb{R}$ by $V_i(z, l) = \sum_{j \in I} z(j) v_{i,l}(j)$ and $\widetilde{V}_i(z, l) = \sum_{j \in I} z(j) \tilde{v}_{i,l}(j)$, respectively. Consider $l'$ for which $\succ_{i,l'} \neq \emptyset$. Since both the functions $v_{i,l'}$ and $\tilde{v}_{i,l'}$ are vN-M representations of $\succcurlyeq_{i,l'}$, it follows that there exists constants $\tau_i > 0$, $\tau_i'$ such that for all $(z, l') \in \Delta(I) \times \{l'\}$, $\widetilde{V}_i(z, l') = \tau_i V_i(z, l') + \tau_i'$. Now, consider any $l \neq l'$. First, consider the case where $\succ_{i,l} = \emptyset$. In this case, we know from Lemma A.1 that there exists $z^* \in \Delta(I)$ such that $(z, l) \sim_i^* (z^*, l')$, for all $z \in \Delta(I)$. Accordingly, $\widetilde{V}_i(z, l) = \widetilde{V}_i(z^*, l') = \tau_i V_i(z^*, l') + \tau_i' = \tau_i V_i(z, l) + \tau_i'$. Next consider $l \neq l'$ such that $\succ_{i,l} \neq \emptyset$. We know from Lemma A.1 that there exists $\widetilde{z}$, $\widehat{z}$, $\widetilde{z}'$, $\widehat{z}' \in \Delta(I)$ such that $(\widetilde{z}, l) \sim_i^* (\widetilde{z}', l') \succ_i^* (\widehat{z}, l) \sim_i^* (\widehat{z}', l')$. Further, we can find constants $\tau_i(l) > 0$ and $\tau_i'(l)$ such that for all $(z, l) \in \Delta(I) \times \{l\}$, $\widetilde{V}_i(z, l) = \tau_i(l) V_i(z, l) + \tau_i'(l)$. Accordingly, it follows that:

$$\begin{aligned} \widetilde{V}_i(\widetilde{z}, l) - \widetilde{V}_i(\widehat{z}, l) &= \widetilde{V}_i(\widetilde{z}', l') - \widetilde{V}_i(\widehat{z}', l') \\ \implies \tau_i(l)[V_i(\widetilde{z}, l) - V_i(\widehat{z}, l)] &= \tau_i[V_i(\widetilde{z}', l') - V_i(\widehat{z}', l')] \end{aligned}$$

Given that $V_i(\widetilde{z}, l) - V_i(\widehat{z}, l) = V_i(\widetilde{z}', l') - V_i(\widehat{z}', l') > 0$, it follows that $\tau_i(l) = \tau_i$. Further, since $\widetilde{V}_i(\widetilde{z}, l) = \widetilde{V}_i(\widetilde{z}', l')$, it follows that $\tau_i V_i(\widetilde{z}, l) + \tau_i'(l) = \tau_i V_i(\widetilde{z}', l') + \tau_i'$. Given that $V_i(\widetilde{z}, l) = V_i(\widetilde{z}', l')$, it follows that $\tau_i'(l) = \tau_i'$. Accordingly, it follows that for any $(z, l) \in \Delta(I) \times \Delta(X)$, $\widetilde{V}_i(z, l) = \tau_i V_i(z, l) + \tau_i'$. In particular, $\tilde{v}_{i,l} = \tau_i v_{i,l} + \tau_i'$, for all $l \in \Delta(X)$. □

**Lemma A.3.** *If $i \in I_1$ satisfies independence over identity lotteries and interpersonal conflict, then for any $l$, $l' \in \Delta(X)$ there exists $l^* \in \Delta(X)$ and $z$, $z'$, $\widetilde{z}$, $\widetilde{z}' \in \Delta(I)$ such that:*

1. *$(z, l^*)$ and $(z', l^*)$ are, respectively, the risk equivalents of $l$ and $l'$ for $i$*

*2.* $([i], l) \sim_i^* (\widetilde{z}, l^*)$ *and* $([i], l') \sim_i^* (\widetilde{z}', l^*)$.

*Proof.* Let $l = [x_1, \alpha_1; \ldots; x_M, \alpha_M]$ and $l' = [x'_1, \alpha'_1; \ldots; x'_N, \alpha'_N]$. Further, let $\overline{l}$, $\underline{l} \in \Delta(X)$ be such that $([i], \overline{l}) \succcurlyeq_i^* ([i], \tilde{l}) \succcurlyeq_i^* ([i], \underline{l})$ for all $\tilde{l} \in \{l, l', [x_1], \ldots, [x_M], [x'_1], \ldots, [x'_N]\}$. First, consider the case when $([i], \overline{l}) \succ_i^* ([i], \underline{l})$. Interpersonal conflict implies that there exists $j \in I$ such that either $([j], \underline{l}) \succcurlyeq_i^* ([i], \overline{l})$ or $([i], \underline{l}) \succcurlyeq_i^* ([j], \overline{l})$. Accordingly, since $\succcurlyeq_i^*$ is continuous and satisfies independence over identity lotteries, it follows that there exists $l^* = \overline{l}$ or $\underline{l}$ and $\widetilde{z}$, $\widetilde{z}'$, $z_m$, $z'_n \in \Delta(I)$, $m = 1, \ldots, M$ and $n = 1, \ldots, N$, such that $([i], l) \sim_i^* (\widetilde{z}, l^*)$, $([i], l') \sim_i^* (\widetilde{z}', l^*)$, $([i], [x_m]) \sim_i^* (z_m, l^*)$, for $m = 1, \ldots, M$, and $([i], [x'_n]) \sim_i^* (z'_n, l^*)$, for $n = 1, \ldots, N$. A similar conclusion, of course, follows also for the case when $([i], \overline{l}) \sim_i^* ([i], \underline{l})$. Accordingly, $(z, l^*) = (\alpha_1 z_1 + \ldots + \alpha_M z_M, l^*)$ and $(z', l^*) = (\alpha'_1 z'_1 + \ldots + \alpha'_N z'_N, l^*)$ are, respectively, risk equivalents of $l$ and $l'$ for $i$. □

## A.2 Proof of Propositions

**Proof of Proposition 3.1.** Let $l$, $l' \in \Delta(X)$. From Lemma A.3, we know that there exists $(z, l^*)$, $(z', l^*) \in \Delta(I) \times \Delta(X)$ that are, respectively, the risk equivalents of $l$ and $l'$ for $i$. Since, $\succcurlyeq_i^*$ is complete, it follows that either $l \succcurlyeq_i^R l'$ or $l' \succcurlyeq_i^R l$. That is, $\succcurlyeq_i^R$ is complete. To see that $\succcurlyeq_i^R$ satisfies the Archimedean continuity condition let $l \succ_i^R l' \succ_i^R l''$. Arguing along similar lines as in the proof of Lemma A.3, we can show that there exists $(z, l^*), (z', l^*), (z'', l^*) \in \Delta(I) \times \Delta(X)$ that are, respectively, the risk equivalents of $l$, $l'$ and $l''$ for $i$. That is, $(z, l^*) \succ_i^* (z', l^*) \succ_i^* (z'', l^*)$. By continuity of $\succcurlyeq_i^*$, it follows that there exists $\overline{\alpha}$ and $\underline{\alpha} \in (0, 1)$ such that $(\overline{\alpha} z + (1 - \overline{\alpha}) z'', l^*) \succ_i^* (z', l^*) \succ_i^* (\underline{\alpha} z + (1 - \underline{\alpha}) z'', l^*)$. Further, it is also straightforward to establish that $(\overline{\alpha} z + (1 - \overline{\alpha}) z'', l^*)$ and $(\underline{\alpha} z + (1 - \underline{\alpha}) z'', l^*)$ are, respectively, the risk equivalents of $\overline{\alpha} l + (1 - \overline{\alpha}) l''$ and $\underline{\alpha} l + (1 - \underline{\alpha}) l''$. Hence, $\overline{\alpha} l + (1 - \overline{\alpha}) l'' \succ_i^R l' \succ_i^R \underline{\alpha} l + (1 - \underline{\alpha}) l''$, which establishes that $\succcurlyeq_i^R$ satisfies the Archimedean continuity condition. Finally, to establish that $\succcurlyeq_i^R$ satisfies the vNM independence condition, let $l, l', l'' \in \Delta(X)$ be such that $l \succ_i^R l'$. Like we argued above, there exists $(z, l^*), (z', l^*), (z'', l^*) \in \Delta(I) \times \Delta(X)$ that are, respectively, the risk equivalents of $l$, $l'$ and $l''$ for $i$. As such, $(z, l^*) \succ_i^* (z', l^*)$. Since $\succcurlyeq_i^*$ satisfies the independence over identity lotteries axiom, it follows that for any $\alpha \in (0, 1]$, we have $(\alpha z + (1 - \alpha) z'', l^*) \succ_i^* (\alpha z' + (1 - \alpha) z'', l^*)$. Further, it is also straightforward to establish that $(\alpha z + (1 - \alpha) z'', l^*)$ and $(\alpha z' + (1 - \alpha) z'', l^*)$ are,

respectively, the risk equivalents of $\alpha l + (1-\alpha)l''$ and $\alpha l' + (1-\alpha)l''$ for $i$. Hence, $\alpha l + (1-\alpha)l'' \succ_i^R \alpha l' + (1-\alpha)l''$, which establishes that $\succcurlyeq_i^R$ satisfies the vNM independence condition.

**Proof of Proposition 3.2.** It follows immediately from Lemma A.3 that $\succcurlyeq_i^P$ is complete. To establish that it is transitive, consider $l, l', l'' \in \Delta(X)$ such that $l \succcurlyeq_i^P l'$ and $l' \succcurlyeq_i^P l''$. We can establish along similar lines as in the proof of Lemma A.3 that there exists $l^* \in \Delta(X)$ and $z$, $z'$, $z''$, $\widetilde{z}$, $\widetilde{z}'$, $\widetilde{z}'' \in \Delta(I)$ such that:

1. $(z, l^*)$, $(z', l^*)$ and $(z'', l^*)$ are, respectively, the risk equivalents of $l, l'$ and $l''$ for $i$
2. $([i], l) \sim_i^* (\widetilde{z}, l^*)$, $([i], l') \sim_i^* (\widetilde{z}', l^*)$ and $([i], l'') \sim_i^* (\widetilde{z}'', l^*)$.

Accordingly,

$$\begin{aligned} l \succcurlyeq_i^P l' &\Rightarrow (.5\widetilde{z} + .5z', l^*) \succcurlyeq_i^* (.5\widetilde{z}' + .5z, l^*) \\ l' \succcurlyeq_i^P l'' &\Rightarrow (.5\widetilde{z}' + .5z'', l^*) \succcurlyeq_i^* (.5\widetilde{z}'' + .5z', l^*) \end{aligned}$$

Since, $\succcurlyeq_i^*$ satisfies independence over identity lotteries, it follows that

$$\begin{aligned} (.5(.5\widetilde{z} + .5z') + .5(.5\widetilde{z}' + .5z''), l^*) &\succcurlyeq_i^* (.5(.5\widetilde{z}' + .5z) + .5(.5\widetilde{z}'' + .5z'), l^*) \\ \Rightarrow (.5(.5\widetilde{z} + .5z'') + .5(.5\widetilde{z}' + .5z'), l^*) &\succcurlyeq_i^* (.5(.5\widetilde{z}'' + .5z) + .5(.5\widetilde{z}' + .5z'), l^*) \\ \Rightarrow (.5\widetilde{z} + .5z'', l^*) &\succcurlyeq_i^* (.5\widetilde{z}'' + .5z, l^*) \end{aligned}$$

Therefore, $l \succcurlyeq_i^P l''$ and $\succcurlyeq_i^P$ is transitive.

**Proof of Proposition 3.3.** Please refer to the proof of Theorem 3.1 below. The proof of this Proposition is established in the course of proving that Theorem.

## A.3 Proof of Theorem 3.1

We now prove Theorem 3.1. So, consider $i \in I_1$ whose preferences satisfy independence over identity lotteries, self-acceptance, interpersonal conflict and revealed fairness.

We first show that $\succcurlyeq_i$ has a basic KS representation. We know that for any such $i \in I_1$, $\succ_i \neq \emptyset$. That is, there exists $l, l' \in \Delta(X)$ such that $l \succ_i l'$. By self-acceptance, we have $([i], l) \succ_i^* ([i], l')$. Interpersonal conflict, then, implies that there exists $j \in I$ such that either $([j], l') \succcurlyeq_i^* ([i], l)$ or $([i], l') \succcurlyeq_i^* ([j], l)$. That is, there exists $l^* = l$ or $l'$ such that $\succ_{i,l^*} \neq \emptyset$. Then, it follows from Lemma A.2 that there exists, for each $l \in \Delta(X)$, a function $v_{i,l} : I \to \mathbb{R}$ such that the function $V_i : \Delta(I) \times \Delta(X) \to \mathbb{R}$ given by $V_i(z, l) = \sum_{j \in I} z(j) v_{i,l}(j)$ represents $\succcurlyeq_i^*$. Define the function $\hat{U}_i : \Delta(X) \to \mathbb{R}$ by $\hat{U}_i(l) = V_i([i], l) = v_{i,l}(i)$. By self-acceptance, $\hat{U}_i$ represents $\succcurlyeq_i$, since:

$$l \succcurlyeq_i l' \iff ([i], l) \succcurlyeq_i^* ([i], l') \iff V_i([i], l) \geq V_i([i], l') \iff \hat{U}_i(l) \geq \hat{U}_i(l')$$

Next, define the functions $u_i : X \to \mathbb{R}$ and $g_i : \Delta(X) \to \mathbb{R}$ that we need to specify as part of a basic KS representation. First, define $u_i$ by $u_i(x) = \hat{U}_i([x]) = V_i([i], [x])$. To define $g_i$, consider any $l = [x_1, \alpha_1; \ldots; x_M, \alpha_M] \in \Delta(X)$. Based on Lemma A.3, we can conclude that there exists exists $l^* \in \Delta(X)$ and $z_m \in \Delta(I)$ such that $([i], [x_m]) \sim_i^* (z_m, l^*)$, for $m = 1, \ldots, M$. That is, $(\alpha_1 z_1 + \ldots + \alpha_M z_M, l^*)$ is a risk equivalent of $l$ for $i$. Define,

$$g_i(l) = V_i([i], l) - V_i(\alpha_1 z_1 + \ldots + \alpha_M z_M, l^*) = V_i([i], l) - \sum_{m=1}^M \alpha_m V_i(z_m, l^*).$$

Accordingly,

$$g_i(l) = \hat{U}_i(l) - \sum_{m=1}^M \alpha_m V_i([i], [x_m]) = \hat{U}_i(l) - \sum_{m=1}^M \alpha_m u_i(x_m).$$

In other words, the function $\hat{U}_i : \Delta(X) \to \mathbb{R}$, given by $\hat{U}_i(l) = \sum_{x \in X} l(x) u_i(x) + g_i(l)$, represents $\succcurlyeq_i$. Observe that, clearly, $g_i([x]) = 0$, for all $x \in X$.

Now, to establish that $(u_i, g_i)$ is indeed a basic KS representation of $\succcurlyeq_i$, we need to show that $(i)$ $u_i$ is a vNM representation of $\succcurlyeq_i^R$; and $(ii)$ $g_i$ represents the binary relation $\succcurlyeq_i^P$ and is a continuous, strictly quasi-concave function. To that end, consider any $l = [x_1, \alpha_1; \ldots, x_M, \alpha_M]$, $l' = [x'_1, \alpha'_1; \ldots, x'_N, \alpha'_N] \in \Delta(X)$. From Lemma A.3, we know that there exists $l^* \in \Delta(X)$ and $z = \alpha_1 z_1 + \cdots + \alpha_M z_M$, $z' = \alpha'_1 z'_1 + \cdots + \alpha'_N z'_N$, $\widetilde{z}$, $\widetilde{z}' \in \Delta(I)$ such that:

1. $(z, l^*)$ and $(z', l^*)$ are, respectively, the risk equivalents of $l$ and $l'$ for $i$
2. $([i], l) \sim_i^* (\widetilde{z}, l^*)$ and $([i], l') \sim_i^* (\widetilde{z}', l^*)$.

Now, to see that $u_i$ is a vNM representation of $\succcurlyeq_i^R$, observe that:

$$\begin{aligned}
l \succcurlyeq_i^R l' &\Leftrightarrow (\alpha_1 z_1 + \cdots + \alpha_M z_M, l^*) \succcurlyeq_i^* (\alpha'_1 z'_1 + \cdots + \alpha'_N z'_N, l^*) \\
&\Leftrightarrow V_i(\alpha_1 z_1 + \cdots + \alpha_M z_M, l^*) \geq V_i(\alpha'_1 z'_1 + \cdots + \alpha'_N z'_N, l^*) \\
&\Leftrightarrow \sum_{m=1}^{M} \alpha_m V_i(z_m, l^*) \geq \sum_{n=1}^{N} \alpha'_n V_i(z'_n, l^*) \\
&\Leftrightarrow \sum_{m=1}^{M} \alpha_m V_i([i], [x_m]) \geq \sum_{n=1}^{N} \alpha'_n V_i([i], [x'_n]) \\
&\Leftrightarrow \sum_{m=1}^{M} \alpha_m u_i(x_m) \geq \sum_{n=1}^{N} \alpha'_n u_i(x'_n)
\end{aligned}$$

Next, to establish that $g_i$ represents $\succcurlyeq_i^P$, observe that:

$$\begin{aligned}
l \succcurlyeq_i^P l' &\Leftrightarrow (.5\widetilde{z} + .5z', l^*) \succcurlyeq_i^* (.5\widetilde{z}' + .5z, l^*) \\
&\Leftrightarrow V_i(.5\widetilde{z} + .5z', l^*) \geq V_i(.5\widetilde{z}' + .5z, l^*) \\
&\Leftrightarrow .5V_i(\widetilde{z}, l^*) + .5V_i(z', l^*) \geq .5V_i(\widetilde{z}', l^*) + .5V_i(z, l^*) \\
&\Leftrightarrow V_i(\widetilde{z}, l^*) + \sum_{n=1}^{N} \alpha'_n V_i(z'_n, l^*) \geq V_i(\widetilde{z}', l^*) + \sum_{m=1}^{M} \alpha_m V_i(z_m, l^*) \\
&\Leftrightarrow V_i([i], l) + \sum_{n=1}^{N} \alpha'_n V_i([i], [x'_n]) \geq V_i([i], l') + \sum_{m=1}^{M} \alpha_m V_i([i], [x_m]) \\
&\Leftrightarrow \hat{U}_i(l) + \sum_{n=1}^{N} \alpha'_n u_i(x'_n) \geq \hat{U}_i(l') + \sum_{m=1}^{M} \alpha_m u_i(x_m) \\
&\Leftrightarrow \hat{U}_i(l) + \hat{U}_i(l') - g_i(l') \geq \hat{U}_i(l') + \hat{U}_i(l) - g_i(l) \\
&\Leftrightarrow g_i(l) \geq g_i(l')
\end{aligned}$$

At this point, note that we have established parts $(i)$ and $(ii)$ of Proposition 3.3.[15]

To establish that $g_i$ is continuous, it suffices to show that $\hat{U}_i$ is continuous. To that end, first, note that, since $\succcurlyeq_i^*$ is continuous and $\Delta(X)$ is a compact set,

[15] Observe that thus far in the proof, we have not made any use of the assumption that $i$'s preferences satisfy revealed fairness.

there exists $\overline{l}$, $\underline{l} \in \Delta(X)$ such that $([i], \overline{l}) \succcurlyeq_i^* ([i], l) \succcurlyeq_i^* ([i], \underline{l})$, for all $l \in \Delta(X)$. Self-acceptance, in turn, implies that $\overline{l} \succcurlyeq_i l \succcurlyeq_i \underline{l}$, for all $l \in \Delta(X)$. Now take any $c \in \mathbb{R}$. To establish that $\hat{U}_i$ is continuous, it is sufficient to show that the sets $\{l \in \Delta(X) : \hat{U}_i(l) \geq c\}$ and $\{l \in \Delta(X) : \hat{U}_i(l) \leq c\}$ are closed sets. If $c > \hat{U}_i(\overline{l})$ or $c < \hat{U}_i(\underline{l})$, then the conclusion is immediate. So, consider $c$ such that $\hat{U}_i(\underline{l}) \leq c \leq \hat{U}_i(\overline{l})$. Let $(z', l') \in \Delta(I) \times \Delta(X)$ be such that $V_i(z', l') = c$. We know that such a $(z', l')$ exists because the range of $V_i$ is connected and the range of $\hat{U}_i$ is a subset of the range of $V_i$. Hence,

$$\begin{aligned} \{l \in \Delta(X) : \hat{U}_i(l) \geq c\} &= \{l \in \Delta(X) : ([i], l) \succcurlyeq_i^* (z', l')\} \\ \{l \in \Delta(X) : \hat{U}_i(l) \leq c\} &= \{l \in \Delta(X) : (z', l') \succcurlyeq_i^* ([i], l)\} \end{aligned}$$

These sets are closed since $\succcurlyeq_i^*$ is continuous.

Finally, we establish that $g_i$ is strictly quasi-concave. So, consider $l$, $l' \in \Delta(X)$ with $l \neq l'$. We need to show that $g_i(\alpha l + (1-\alpha)l') > \min\{g_i(l),\ g_i(l')\}$, for all $\alpha \in (0,1)$. First, suppose $g_i(l) = g_i(l')$. Since $g_i$ represents $\succcurlyeq_i^P$, it follows that $l \sim_i^P l'$. Revealed fairness then implies that for all $\alpha \in (0,1)$, $\alpha l + (1-\alpha)l' \succ_i^P l$, which, in turn, implies that $g_i(\alpha l + (1-\alpha)l') > g_i(l) = \min\{g_i(l), g_i(l')\}$. Next, consider the case when $g_i(l) \neq g_i(l')$—w.l.o.g. suppose, $g_i(l) > g_i(l')$. Suppose, towards a contradiction, that there exists $\hat{\alpha} \in (0,1)$ such that $g_i(l') \geq g_i(\hat{\alpha} l + (1-\hat{\alpha})l')$. First, consider the possibility that $g_i(l') > g_i(\hat{\alpha} l + (1-\hat{\alpha})l')$. Define the function $f_i : [0,1] \to \mathbb{R}$ by $f_i(\beta) = g_i(\beta l + (1-\beta)(\hat{\alpha} l + (1-\hat{\alpha})l'))$. Given that $g_i$ is continuous, so is $f_i$. It then follows from the intermediate value theorem that since $f_i(0) = g_i(\hat{\alpha} l + (1-\hat{\alpha})l') < g_i(l') < g_i(l) = f_i(1)$, there exists $\hat{\beta} \in (0,1)$ such that $f_i(\hat{\beta}) = g_i(\hat{\beta} l + (1-\hat{\beta})(\hat{\alpha} l + (1-\hat{\alpha})l')) = g_i(l')$. That is, $(\hat{\beta}(1-\hat{\alpha}) + \hat{\alpha})l + (1 - (\hat{\beta}(1-\hat{\alpha}) + \hat{\alpha}))l' \sim_i^P l'$. Revealed fairness then implies that $\hat{\alpha} l + (1-\hat{\alpha})l' \succ_i^P l'$.[16] That is, $g_i(\hat{\alpha} l + (1-\hat{\alpha})l') > g_i(l')$, which brings us to our desired contradiction. Next, consider the possibility that $g_i(l') = g_i(\hat{\alpha} l + (1-\hat{\alpha})l')$. Define, $\tilde{l} = 0.5l' + 0.5(\hat{\alpha} l + (1-\hat{\alpha})l')$. It follows from revealed fairness and the fact that $g_i$ represents $\succcurlyeq_i^P$ that $g_i(\tilde{l}) > g_i(\hat{\alpha} l + (1-\hat{\alpha})l')$. In this case too, following a similar argument as above, we arrive at a contradiction. Therefore, $g_i$ is strictly quasi-concave. This also helps to establish the final part of Proposition 3.3.

All of this together establishes that $(u_i, g_i)$ is a basic KS representation of $\succcurlyeq_i$. Drawing on it, we proceed to show that every representation of $\succcurlyeq_i$ is a KS repre-

[16] To see this, not that by revealed fairness, $\alpha[(\hat{\beta}(1-\hat{\alpha})+\hat{\alpha})l + (1-(\hat{\beta}(1-\hat{\alpha})+\hat{\alpha}))l'] + (1-\alpha)l' \succ_i^P l'$, for all $\alpha \in (0,1)$. The desired conclusion follows by taking $\alpha = \frac{\hat{\alpha}}{\hat{\beta}(1-\hat{\alpha})+\hat{\alpha}}$.

sentation. So, consider any representation $U_i : \Delta(X) \to \mathbb{R}$ of $\succcurlyeq_i$. Define the function $\psi_i : \{(\mathbb{E}u_i(l), g_i(l)) : l \in \Delta(X)\} \to \mathbb{R}$ by $\psi_i(\mathbb{E}u_i(l), g_i(l)) = U_i(l)$, for any $l \in \Delta(X)$. Observe that the function $\psi_i$ is well-defined. To see this, consider $l$ and $l'$ such that $\mathbb{E}u_i(l) = \mathbb{E}u_i(l')$ and $g_i(l) = g_i(l')$. In this case, $\hat{U}_i(l) = \hat{U}_i(l')$ and, so, $l \sim_i l'$. Accordingly, $U_i(l) = U_i(l')$. Finally, we establish that $\psi_i$ is increasing in both its arguments. So, consider $l$ and $l'$ such that $\mathbb{E}u_i(l) > \mathbb{E}u_i(l')$ and $g_i(l) = g_i(l')$. In this case, $\hat{U}_i(l) > \hat{U}_i(l')$, i.e., $l \succ_i l'$, and, so, $U_i(l) = \psi_i(\mathbb{E}u_i(l), g_i(l)) > \psi_i(\mathbb{E}u_i(l'), g_i(l')) = U_i(l')$. This establishes that $\psi_i$ is increasing in its first argument. A similar argument establishes that it is also increasing in its second argument.

## A.4 Proof of Theorem 4.1

We first prove the sufficiency of the axioms for the collection of ethical preferences $(\succcurlyeq_i^*)_{i \in I}$ to admit generalized utilitarian representations that incorporate individuals' sense of justice. So, assume that each $j \in I$ satisfies independence over identity lotteries, the acceptance principle and interpersonal conflict; and, in addition, each $j \in I_1$ satisfies revealed fairness. As pointed out earlier, if any $j$ satisfies the acceptance principle, then he satisfies self-acceptance. Therefore, based on Theorem 3.1, we know that for each $j \in I_1$ there exists a KS representation $(u_j, g_j, \psi_j)$ of $\succcurlyeq_j$. That is, for each $j \in I_1$, there exist functions $u_j : X \to \mathbb{R}$, $g_j : \Delta(X) \to \mathbb{R}$ that is continuous and strictly quasiconcave with $g_j([x]) = 0$ for any $x \in X$, and $\psi_j : \{(\mathbb{E}u_j(l), g_j(l)) : l \in \Delta(X)\} \to \mathbb{R}$ that is increasing in both its arguments, such that the function $U_j : \Delta(X) \to \mathbb{R}$ given by

$$U_j(l) = \psi_j(\mathbb{E}u_j(l), g_j(l))$$

represents $\succcurlyeq_j$. Further, for each $j \in I_0$, there exists a vNM representation of $\succcurlyeq_j$. That is, for each $j \in I_0$, there exists a function $u_j : X \to \mathbb{R}$, such that the function $U_j : \Delta(X) \to \mathbb{R}$ given by $U_j(l) = \mathbb{E}u_j(l)$ represents $\succcurlyeq_j$.

Now, consider a particular $i \in I$ and recall our assumption that there exists some $j \in I$ for whom $\succ_j \neq \emptyset$. Accordingly, by the acceptance principle, it follows that $([j], l') \succ_i^* ([j], l'')$, for some $l', l'' \in \Delta(X)$. Interpersonal conflict then implies that there exists $k \in I$ such that either $([k], l'') \succcurlyeq_i^* ([j], l')$ or $([j], l'') \succcurlyeq_i^* ([k], l')$. That

is, there exists $l^* = l'$ or $l''$ such that $\succ_{i,l^*} \neq \emptyset$. As such, Lemma A.2 implies that for each $l \in \Delta(X)$, there exists a function $v_{i,l} : I \to \mathbb{R}$ such that the function $V_i : \Delta(I) \times \Delta(X) \to \mathbb{R}$ given by

$$V_i(z, l) = \sum_{j \in I} z(j) v_{i,l}(j)$$

represents $\succcurlyeq_i^*$. Further, if $(\tilde{v}_{i,l})_{l \in \Delta(X)}$ is another family of functions that represents $\succcurlyeq_i^*$ in this sense, then there exists constants $\tau_i > 0$ and $\tau_i'$ such that $\tilde{v}_{i,l} = \tau_i v_{i,l} + \tau_i'$, for all $l \in \Delta(X)$.

Next, for each $j \in I$, define the function $U_{i,j} : \Delta(X) \to \mathbb{R}$ given by, $U_{i,j}(l) = v_{i,l}(j)$. By the acceptance principle, it follows that the function $U_{i,j}$ represents $\succcurlyeq_j$, since

$$\begin{aligned} l \succcurlyeq_j l' &\iff ([j], l) \succcurlyeq_i^* ([j], l') \iff V_i([j], l) \geq V_i([j], l') \\ &\iff v_{i,l}(j) \geq v_{i,l'}(j) \iff U_{i,j}(l) \geq U_{i,j}(l') \end{aligned}$$

Accordingly, there exists a monotone function $\phi_{i,j} : \mathbb{R} \to \mathbb{R}$, such that for each $l \in \Delta(X)$, $U_{i,j}(l) = \phi_{i,j}(U_j(l))$. Therefore,

$$\begin{aligned} V_i(z, l) &= \sum_{j \in I} z(j) v_{i,l}(j) = \sum_{j \in I} z(j) U_{i,j}(l) \\ &= \sum_{j \in I} z(j) \phi_{i,j}(U_j(l)) \end{aligned}$$

Hence, the collection of ethical preference relations $(\succcurlyeq_i^*)_{i \in I}$ have generalized utilitarian representations $((u_i)_{i \in I_0}, (u_i, g_i, \psi_i)_{i \in I_1}, ((\phi_{ij})_{j \in I})_{i \in I})$ that incorporate individuals' sense of justice. It is straightforward to establish that if such representations exist then each $i \in I$ satisfies independence over identity lotteries and the acceptance principle; and each $i \in I_1$ satisfies revealed fairness. We omit the details here.

The second part of the theorem establishing the (essential) uniqueness properties of such representations follows in a straightforward way from the second part of Lemma A.2. We omit those details as well.

## A.5 Proof of Theorem 4.2

We prove the sufficiency of the axioms for the collection of ethical preferences $(\succcurlyeq_i^*)_{i \in I}$ to admit utilitarian representations that incorporate individuals' sense of

justice. So, assume that each $j \in I$ satisfies independence over identity lotteries, self-acceptance and interpersonal conflict; each $j, k \in I$ satisfies shared ethics; and, in addition, each $j \in I_1$ satisfies revealed fairness and each $j \in I_0$ satisfies indifference to similar randomizations.

Accordingly, based on Theorem 3.1, we know that for each $j \in I_1$, there exists a basic KS representation $(u_j, g_j)$ of $\succcurlyeq_j$. That is, for each $j \in I_1$, there exist functions $u_j : X \to \mathbb{R}$ and $g_j : \Delta(X) \to \mathbb{R}$ that is continuous and strictly quasiconcave with $g_j([x]) = 0$ for any $x \in X$, such that the function $U_j : \Delta(X) \to \mathbb{R}$, given by

$$U_j(l) = \mathbb{E}u_j(l) + g_j(l),$$

represents $\succcurlyeq_j$. Further, for each $j \in I_0$, there exists a vNM representation of $\succcurlyeq_j$. That is, for each $j \in I_0$, there exists a function $u_j : X \to \mathbb{R}$, such that the function $U_j : \Delta(X) \to \mathbb{R}$ given by $U_j(l) = \mathbb{E}u_j(l)$ represents $\succcurlyeq_j$.

Next, recall that there exists $j' \in I$ such that $\succ_{j'} \neq \emptyset$. That is, there exists $\tilde{l}$, $\hat{l} \in \Delta(X)$ such that $\tilde{l} \succ_{j'} \hat{l}$. By self acceptance, it follows that $([j'], \tilde{l}) \succ^*_{j'} ([j'], \hat{l})$. As we have seen above, interpersonal conflict, then, implies that there exists $l' = \tilde{l}$ or $\hat{l}$, such that $\succ_{j',l'} \neq \emptyset$. By shared ethics, this implies that $\succ_{j,l'} \neq \emptyset$ for any $j \in I$. Accordingly, Lemma A.2 implies that for any such $j$, there exists, for all $l \in \Delta(X)$, a function $v_{j,l} : I \to \mathbb{R}$ such that the function $V_j : \Delta(I) \times \Delta(X) \to \mathbb{R}$ given by

$$V_j(z, l) = \sum_{k \in I} z(k) v_{j,l}(k) = \sum_{k \in I} z(k) V_j([k], l)$$

represents $\succcurlyeq^*_j$. Further, if $(\tilde{v}_{j,l})_{l \in \Delta(X)}$ is another family of functions that represents $\succcurlyeq^*_j$ in this sense, then there exists constants $\tau_j > 0$ and $\tau'_j$ such that $\tilde{v}_{j,l} = \tau_j v_{j,l} + \tau'_j$, for all $l \in \Delta(X)$.

Now, fix $i \in I$ and consider any $j \neq i$. We next establish that for any such $j$ we can re-calibrate the $V_j$ function derived above to ensure that $V_j(z, l) = V_i(z, l)$ for any $(z, l) \in \Delta(I) \times \Delta(X)$. To that end, as shown above, note that there exists $l' \in \Delta(X)$ such that $\succ_{i,l'} \neq \emptyset$ and $\succ_{j,l'} \neq \emptyset$. That is, by shared ethics, there exists $z', z'' \in \Delta(I)$ such that $(z', l') \succ^*_i (z'', l')$ and $(z', l') \succ^*_j (z'', l')$. We know from the proof of Lemma A.2 that the function $V_j$ is defined uniquely up to a positive affine transformation, i.e., we have two degrees of freedom in terms of defining it. Re-normalize $V_j$ by setting $V_j(z', l') = V_i(z', l')$ and $V_j(z'', l') = V_i(z'', l')$.

It is now easy to establish that for any $z \in \Delta(I)$, $V_j(z,l') = V_i(z,l')$. To see this, first, consider $z \in \Delta(I)$ such that $(z',l') \succcurlyeq_i^* (z,l') \succcurlyeq_i^* (z'',l')$, so that, by shared ethics, $(z',l') \succcurlyeq_j^* (z,l') \succcurlyeq_j^* (z'',l')$. This same axiom in conjunction with independence over identity lotteries and continuity of ethical preferences, then, implies that there exists a unique $\alpha \in [0,1]$ such that $(z,l') \sim_i^* (\alpha z' + (1-\alpha)z'',l')$ and $(z,l') \sim_j^* (\alpha z' + (1-\alpha)z'',l')$. Hence,

$$\begin{aligned} V_i(z,l') &= V_i(\alpha z' + (1-\alpha)z'',l') \\ &= \alpha V_i(z',l') + (1-\alpha)V_i(z'',l') = \alpha V_j(z',l') + (1-\alpha)V_j(z'',l') \\ &= V_j(\alpha z' + (1-\alpha)z'',l') = V_j(z,l') \end{aligned}$$

We can establish along similar lines that $V_i(z,l') = V_j(z,l')$ for $z \in \Delta(I)$ such that $(z,l') \succ_i^* (z',l')$ or $(z'',l') \succ_i^* (z,l')$.

We now extend the above conclusion by showing that, in fact, $V_j(z,l) = V_i(z,l)$, for any $(z,l) \in \Delta(I) \times \Delta(X)$. To that end, first, consider those $l \in \Delta(X)$ for which $\succ_{i,l} = \emptyset$ and, hence, by shared ethics, $\succ_{j,l} = \emptyset$. We know from lemma A.1 that, in this case, there exists $\widetilde{z} \in \Delta(I)$ such that $(\widetilde{z},l') \sim_i^* (z,l)$, for all $z \in \Delta(I)$. By shared ethics, it follows that $(\widetilde{z},l') \sim_j^* (z,l)$, for all $z \in \Delta(I)$. Based on our conclusion above, it follows that

$$V_i(z,l) = V_i(\widetilde{z},l') = V_j(\widetilde{z},l') = V_j(z,l)$$

Next, consider those $l \in \Delta(X)$ for which $\succ_{i,l} \neq \emptyset$ and, hence, $\succ_{j,l} \neq \emptyset$. We know from lemma A.1 that, in this case, there exists $\widetilde{z}$, $\widehat{z}$, $\widetilde{z}'$, $\widehat{z}' \in \Delta(I)$ such that $(\widetilde{z},l) \sim_i^* (\widetilde{z}',l') \succ_i^* (\widehat{z},l) \sim_i^* (\widehat{z}',l')$, and, hence, $(\widetilde{z},l) \sim_j^* (\widetilde{z}',l') \succ_j^* (\widehat{z},l) \sim_j^* (\widehat{z}',l')$. It then follows that:

$$\begin{aligned} V_i(\widetilde{z},l) = V_i(\widetilde{z}',l') &= V_j(\widetilde{z}',l') = V_j(\widetilde{z},l), \text{ and} \\ V_i(\widehat{z},l) = V_i(\widehat{z}',l') &= V_j(\widehat{z}',l') = V_j(\widehat{z},l) \end{aligned}$$

So, since $V_i(\widetilde{z},l) = V_j(\widetilde{z},l) > V_i(\widehat{z},l) = V_j(\widehat{z},l)$, we can establish along similar lines as above that $V_i(z,l) = V_j(z,l)$ for all $z \in \Delta(I)$. We have therefore reached our desired conclusion that $V_j(z,l) = V_i(z,l)$ for all $(z,l) \in \Delta(I) \times \Delta(X)$.

Next, we show that for any $j \in I$ we can re-calibrate the function $U_j$ identified above to ensure that for all $l \in \Delta(X)$, $v_{j,l}(j) = V_j([j],l) = U_j(l)$. This conclusion

has already been established for any $j \in I_1$ in the course of proving Theorem 3.1. In that proof, we showed that the functions $u_j$ and $g_j$ can be defined in such a way that for all $l \in \Delta(X)$, $v_{j,l}(j) = V_j([j], l) = U_j(l)$. Therefore, what remains to be shown is that this conclusion can also be established for any $j \in I_0$. So, consider any $j \in I_0$. If $\succ_j = \emptyset$, then this is straightforward to establish. Simply set the constant $u_j$ function equal to $v_{j,l}(j)$ for some $l \in \Delta(X)$. This is well defined, since by self acceptance $v_{j,l}(j) = v_{j,l'}(j)$, for all $l, l' \in \Delta(X)$. Now consider the case when $\succ_j \neq \emptyset$. Let $\bar{l}$ and $\underline{l}$ be such that $\bar{l} \succcurlyeq_j l \succcurlyeq_j \underline{l}$ for all $l \in \Delta(X)$. We established while proving Theorem 3.1 that such $\bar{l}$ and $\underline{l}$ exist owing to continuity of preferences. We know that there are two degrees of freedom in specifying the function $u_j$ and, accordingly, $U_j$. Re-calibrate the function $U_j$ (and correspondingly $u_j$) by setting $U_j(\bar{l}) = V_j([j], \bar{l})$ and $U_j(\underline{l}) = V_j([j], \underline{l})$. We now show that $U_j(l) = V_j([j], l)$, for all $l \in \Delta(X)$. To that end, first note that interpersonal conflict along with independence over identity lotteries and continuity of ethical preferences imply that there exists $\bar{z}, \underline{z} \in \Delta(I)$ and $l^* \in \Delta(X)$ such that $([j], \bar{l}) \sim_j^* (\bar{z}, l^*)$ and $([j], \underline{l}) \sim_j^* (\underline{z}, l^*)$. Consider any $l \in \Delta(X)$. By virtue of the fact that $\succcurlyeq_j$ is vNM, it follows that there exists a unique $\beta \in [0, 1]$ such that $l \sim_j \beta\bar{l} + (1-\beta)\underline{l}$ and, hence, $U_j(l) = U_j(\beta\bar{l} + (1-\beta)\underline{l})$. Self-acceptance implies that $([j], l) \sim_j^* ([j], \beta\bar{l} + (1-\beta)\underline{l})$. Further, indifference to similar randomizations implies that $([j], \beta\bar{l} + (1-\beta)\underline{l}) \sim_j^* (\beta\bar{z} + (1-\beta)\underline{z}, l^*)$. Accordingly,

$$\begin{aligned} V_j([j], l) &= V_j([j], \beta\bar{l} + (1-\beta)\underline{l}) = V_j(\beta\bar{z} + (1-\beta)\underline{z}, l^*) \\ &= \beta V_j(\bar{z}, l^*) + (1-\beta) V_j(\underline{z}, l^*) = \beta V_j([j], \bar{l}) + (1-\beta) V_j([j], \underline{l}) \\ &= \beta U_j(\bar{l}) + (1-\beta) U_j(\underline{l}) = U_j(\beta\bar{l} + (1-\beta)\underline{l}) = U_j(l) \end{aligned}$$

Therefore, we have established that for all $j \in I$ and all $l \in \Delta(X)$, $v_{j,l}(j) = V_j([j], l) = U_j(l)$.

Now, consider any $i \in I$. Putting everything together, we have that

$$\begin{aligned} V_i(z, l) &= \sum_{j \in I} z(j) v_{i,l}(j) = z(i) v_{i,l}(i) + \sum_{j \neq i} z(j) v_{i,l}(j) \\ &= z(i) V_i([i], l) + \sum_{j \neq i} z(j) V_i([j], l) = z(i) U_i(l) + \sum_{j \neq i} z(j) V_j([j], l) \\ &= z(i) U_i(l) + \sum_{j \neq i} z(j) U_j(l) \\ &= \sum_{j \in I} z(j) U_j(l) = V_k(z, l) = V(z, l) \end{aligned}$$

This completes the proof of the sufficiency of the axioms for the representation. Necessity of the axioms is straightforward to establish, as is the uniqueness result in the second part of the theorem (which follows in a straightforward way from the uniqueness result in Lemma A.2). We omit the details here.

# References


Bolton, G. E., and A. Ockenfels (2000): "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, 90, 166–193.

Brock, J. M., A. Lange, and E. Y. Ozbay (2013): "Dictating the Risk: Experimental Evidence on Giving in Risky Environments," *American Economic Review*, 103, 415–437.

Broome, J. (1984): "Uncertainty and Fairness," *Economic Journal*, 94, 624–632.

——— (1991): *Weighing Goods*, Cambridge: Basil Blackwell.

Charness, G., and M. Rabin (2002): "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics*, 117, 817–869.

Diamond, P. (1967): "Cardinal Welfare, Individualistic Ethics and Interpersonal Comparison of Utility: Comment," *Journal of Political Economy*, 75, 765–766.

Epstein, L. G., and U. Segal (1992): "Quadratic Social Welfare Functions," *Journal of Political Economy*, 100, 691–712.

Fehr, E., and K. M. Schmidt (1999): "A Theory of Fairness, Competition and Cooperation," *Quarterly Journal of Economics*, 114, 817–868.

Fudenberg, D., and D. K. Levine (2012): "Fairness, Risk Preferences and Independence: Impossibility Theorems," *Journal of Economic Behavior and Organization*, 81, 602–612.

Grant, S., A. Kajii, B. Polak, and Z. Safra (2010): "Generalized Utilitarianism and Harsanyi's Impartial Observer Theorem," *Econometrica*, 78, 1939–1971.

Harsanyi, J. (1953): "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking," *Journal of Political Economy*, 61, 434–435.

——— (1955): "Cardinal Welfare, Individualistic Ethics and Interpersonal Comparison of Utility," *Journal of Political Economy*, 63, 309–321.

Karni, E. (1996): "Social Welfare Functions and Fairness," *Social Choice and Welfare*, 13, 487–496.

Karni, E., and Z. Safra (2000): "An extension of a theorem of von Neumann and Morgenstern with an application to social choice theory," *Journal of Mathematical Economics*, 34(3), 315–327.

——— (2002): "Individual Sense of Justice: A Utility Representation," *Econometrica*, 70, 263–284.

Krawczyk, M., and F. LeLec (2010): "Give me a chance! An Experiment in Social Decision under Risk," *Experimental Economics*, 13, 500–511.

Saito, K. (2013): "Social Preferences Under Risk: Equality of Opportunity versus Equality of Outcome," *American Economic Review*, 103, 3084–3101.

Trautmann, S. T. (2010): "Individual Fairness in Harsanyi's Utilitarianism: Operationalizing All-Inclusive Utility," *Theory and Decision*, 68, 405–415.