Welfare vs. Utility

Franz Dietrich
Paris School of Economics & CNRS

Nov. 2022 / Feb. 2026[1]

forthcoming in *Economic Theory*

Abstract

Economists routinely measure individual welfare by (von-Neumann-Morgenstern) utility, for instance when analysing welfare intensity, social welfare, or welfare inequality. Is this welfare measure justified? Natural working hypotheses turn out to imply a different measure. It overcomes familiar problems of utility, by faithfully capturing non-ordinal information, such as welfare intensity – despite still resting on purely ordinal evidence, such as revealed preferences or self-reported welfare comparisons. Social welfare analysis changes when based on this new individual welfare measure rather than utility. For instance, Harsanyi's 'utilitarian theorem' now supports prioritarianism. We compare the standard utility-based versions of utilitarianism and prioritarianism with new versions based on our welfare measure. We show that utility is a hybrid object determined by two rival influences: welfare and the attitude to intrinsic risk, i.e., to risk in welfare. A new version of Harsanyi's theorem shows that Harsanyi makes the questionable implicit assumption that society is neutral to intrinsic risk, overruling people's risk attitudes. We thus propose *risk-impartial utilitarianism*, which adopts people's (average) risk attitude.

Keywords: welfare, utility, risk attitude, social welfare, utilitarianism, Harsanyi-Sen debate, Harsanyi's Theorem, Bernoulli's Hypothesis

1 Introduction

How should someone's welfare be measured? This long-standing question matters in social science and ethics alike. In practice, economists often measure welfare by VNM

[1]This paper has benefited considerably from exchanges with various colleagues, including Matthew Adler, Jean Baccelli, Krister Bykvist, Pietro Cibinel, Marc Fleurbaey, Christian List, Marcus Pivato, Martin Rechenauer, John Weymark, Stéphqne Zuber, and two very helpful anonymous referees. Earlier versions have been presented in various economical or philosophical seminars since 2023, and at the conferences SEAT (Paris, July 2023), Foundations of Decision Theory (LMU Munich, June 2024), Economics & Philosophy Workshop (Lyon Saint-Etienne, March 2025), Frontiers of Economics and Philosophy (Paris School of Economics, May 2025), and PPE Meeting (King's College, London, July 2025). Many thanks to the audiences for interesting feedback and challenges.

utility, largely because VNM utility can be derived from ordinal evidence, such as revealed preferences or self-reported comparisons between risky options. The notorious objection against this welfare measure is that it fails to faithfully reflect non-ordinal welfare features, such as welfare intensity, regardless of how one normalises VNM utility. Non-ordinal welfare features are however indispensable for many applications, such as: aggregating individual into social welfare, comparing welfare levels (or differences) across people, measuring inequality in welfare, and making welfare-based policy recommendations.

The controversy over whether VNM utility can measure welfare has culminated in the Harsanyi-Sen debate in the 1970s (later joined by Weymark 1991, 2005 and others) and counts today among the most persistent points of divergence among welfare economists or formal ethicists. Critics of VNM utility as a welfare measure have so far a difficult standing, as they have not yet come up with an alternative measure based on ordinal evidence. The lack of ordinal foundations exposes these critics to the non-observability objection – at least if one follows the ordinalist tradition that accepts only ordinal evidence. Ordinalism about evidence is itself controversial, but we will stick to this common assumption here (cf. Baccelli and Mongin 2016 on ordinalism and utility).

This paper provides a proof of concept for the VNM-sceptic position, by showing that purely ordinal evidence leads to a *different* welfare measure if one accepts some plausible working assumptions. This measure will respond to classic objections against VNM utility as a welfare measure. We will also take first steps towards a *social* welfare analysis based on that improved measure of individual welfare.

We adopt the familiar idea, defended by Bell and Raiffa (1988) and Cibinel (2025a, 2025b), that someone's VNM utility function is influenced by two distinct features: her welfare or 'intrinsic utility' from outcomes, and her attitude to risk *in welfare* or *intrinsic risk*. Unfortunately, one and the same VNM utility function can be given many different explanations in terms of welfare and intrinsic risk attitude – which *seems* to make welfare unobservable. Figure 1 indeed displays four rival explanations of the same concave VNM

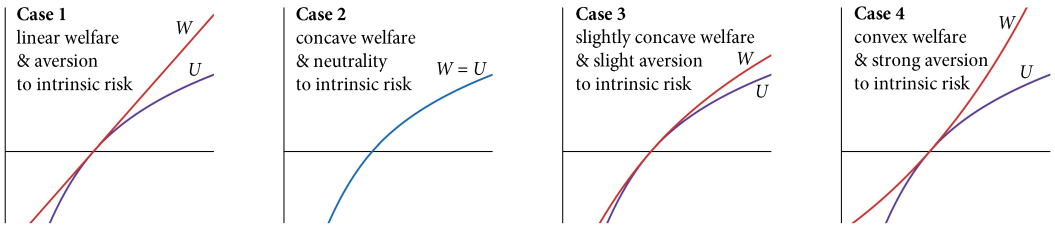

Figure 1: Four explanations of the same VNM utility function (over wealth for instance) in terms of welfare and intrinsic risk attitude

utility function over wealth levels:[2]

[2] In all four plots in Figure 1, the utility and welfare functions are normalised so that they take value 0 and derivative 1 at a fixed wealth level.

- *Case 1: constant marginal welfare & aversion to intrinsic risk.* An extra 1$ gives the same extra welfare regardless of initial wealth; so welfare is linear. Meanwhile a lottery with expected welfare w is worse than a sure outcome with welfare w; so utility is concave *in welfare*, hence also concave in wealth, as welfare is linear. Note that we have invoked the intrinsic rather than standard risk attitude, by considering expected welfare, not expected wealth.
- *Case 2: diminishing marginal welfare & neutrality to intrinsic risk.* An extra \$1 gives less extra welfare if initial wealth is higher; so welfare is concave. Meanwhile a lottery with expected welfare w is as good as a sure welfare of w; so utility coincides with welfare, hence is also concave.
- *Case 3: slightly diminishing marginal welfare & slight aversion to intrinsic risk.* An extra \$1 gives slightly less extra welfare if initial wealth is higher; so welfare is slightly concave. Meanwhile a lottery with expected welfare w is slightly worse than a sure welfare of w; so utility is slightly concave in welfare, and thus concave in wealth.
- *Case 4: increasing marginal welfare & strong aversion to intrinsic risk.* An extra \$1 gives more extra welfare if initial wealth is higher; so welfare is convex. Meanwhile a lottery with expected welfare w is much worse than a sure welfare of w; so utility is so strongly concave in welfare that it is also concave in wealth.

Seemingly, none of the four explanations can be ruled out empirically. This would block the access to personal welfare, and thus block the possibility to assess social welfare, measure welfare inequality, or make welfare-based policy recommendations. It is thus understandable, perhaps, that welfare economists have been naturally drawn towards VNM utility as a proxy for welfare, at least when observability matters. However, the more the persons depart from intrinsic risk neutrality, the less accurate this welfare proxy becomes for them – ultimately distorting social welfare judgments and welfare-based policies.

This paper introduces a new approach by which welfare *will* become observable, without drawing on non-ordinal evidence. But first, what does the literature say?

Welfare and utility in the literature

Our view that VNM utility has two determinants – welfare (intrinsic utility) and attitude to risk in welfare (intrinsic risk attitude) – will resonate with many rational choice theorists, who routinely invoke one or the other or both determinants. Still, this picture can be challenged. We now sketch some prominent understandings of 'utility' and 'welfare', not all of which are compatible with our approach.

Sen (1977) and Weymark (1991, 2005) forcefully distinguish someone's welfare from her VNM utility. For them, someone's VNM utility functions have no privileged status: there are many other functions representing her order, some of which might better capture her welfare in a substantive, more-than-ordinal sense. We agree – and will add a concrete

proposal of a welfare function, which will indeed not be of VNM-type.

Bell and Raiffa (1988), Nissan-Rosen (2015), Dietrich and Jabarian (2022) and many others endorse the distinction between welfare and VNM utility. Yet for instance Harsanyi, Broome (1991), Greaves (2017), and McCarthy et al. (2020) question or even reject the distinction. Fleurbaey and Mongin (2016) take a nuanced view.

For John Harsanyi, welfare just *is* VNM utility. VNM utility correctly quantifies welfare, once it is suitably normalised. Levels and differences of VNM utility are perfectly meaningful quantities that represent how good a situation or change of situation is for the agent. Although this thought is most directly associated with Harsanyi, some of his followers have provided more systematic arguments for the VNM-based account of welfare – perhaps most influentially John Broome, who calls the identification of well-being with VNM utility *Bernoulli's hypothesis*. Like us, Harsanyi takes purely ordinal evidence (a VNM order) to generate a non-ordinal measure of welfare – but, unlike us, he takes VNM utility to be this measure. Again unlike us, he does not distinguish between risk-attitudinal and welfare-based determinants of VNM utility: he treats risk aversion as a mere by-product of diminishing marginal welfare, not as a separate psychological disposition.

Against this reductive view, Bell and Raiffa (1988) restore the conceptual independence between someone's risk-attitudinal and welfare-based features: diminishing marginal welfare is perfectly compatible with 'loving risk', as is increasing marginal welfare with 'hating risk'. We agree, and would add that Harsanyi's conflation of the two phenomena comes from cashing out risk attitudes as attitudes to risk *in outcomes* rather than *in welfare*. This 'non-intrinsic' sort of risk attitude is indeed influenced by (marginal) welfare – but it is not *reducible* to welfare, since it is, like VNM utility, shaped by the combination of welfare and intrinsic risk attitude, as will be shown later.

First-generation economists, such as Gossen, Jevons, Menger, Walras, and Marshall, used to think of utility in ways perfectly independent of any risk or risk attitude. Their term 'utility' corresponds to our 'welfare' or 'intrinsic utility', not to 'VNM utility'. Accordingly, their 'law of diminishing marginal utility' (from consumption) does not reflect risk aversion, but diminishing marginal *welfare or value*. They leave open how to measure welfare from ordinal data – our central problem.

Arrow (1965) and Pratt (1964), the fathers of the modern theory of risk aversion, focus on VNM utility rather than welfare. They take someone's VNM utility function to be shaped entirely and solely by her risk attitude. At first, this seems to deny the idea that VNM utility has two determinants, welfare (intrinsic utility) and risk attitude. This impression however overlooks that Arrow-Pratt's 'risk attitude' is not the intrinsic risk attitude, but a hybrid object influenced by the intrinsic risk attitude and welfare, just like VNM utility. So, while their one-sided labels 'risk attitude' and 'theory of risk aversion' has caused some conceptual confusion by suppressing the role of welfare while stressing 'gambling taste', their theory stands in no formal conflict with our analysis. In passing, we will answer a question that they leave open: how can their measure of (non-intrinsic) risk aversion be reduced to its implicit determinants, intrinsic risk aversion and welfare?

On classical risk attitudes, see also Baccelli (2018).

Peter Wakker (2010) takes an unorthodox, prospect-theoretic view, also endorsed by Abdellaoui et al. (2007) and Buchak (2013). This interesting view separates cleanly between welfare and risk attitude: while welfare affects VNM utility, the risk attitude only affects the weighting of probabilities, within a rank-dependent expected-utility model. Utility is thus explained by welfare alone – but not because the risk attitude is a mere by-product of marginal welfare (as for Harsanyi) but because this attitude leaves its mark elsewhere than in utility. By treating utility as a purely welfare-theoretic, not risk-attitudinal construct, the view conflicts with our two-determinant view on VNM utility, and trivialises our question of how to measure welfare – VNM utility itself is a measure. The view also conflicts with Arrow-Pratt's theory, which links VNM utility to the risk attitude – except under a radical reinterpretation of this theory, as a theory 'of diminishing marginal welfare' rather than 'of risk aversion'.

The measurement of welfare has also been studied extensively in *measurement theory*, under names such as 'measuring strength-of-preference' or 'measuring preference-intensity'. See in particular Krantz et al. (1971), Shapley (1975), Basu (1982), Adler (2016), Nebel (2023, 2024b), and, for particularly general results, Wakker (1988, 1989), Köbberling (2006) and Pivato (2013). This literature pursues an interestingly different agenda, as the evidential basis is not simply an order over alternatives, but some richer structure that is supposed to be represented numerically. That structure is often an order R over alternative *pairs*, called a 'difference order', where $(x, y)R(x', y')$ means 'a change from x to y is at least as good for the person as a change from x' to y''. In this case, one looks for a welfare measure W that represents the difference order.[3] Like a VNM utility function, such a welfare measure is typically unique up to increasing affine transformation. But, unlike a VNM function, it captures welfare *intensity*, not just welfare comparisons. Its evidential basis – the difference order R – is not revealed by choice. It could capture evidence of a different sort, such as information reported by the person, or views of a social planner about her well-being (as in Adler 2025, ch. 4). By contrast, we aim to measure welfare based on ordinal evidence, where 'ordinal' refers to a binary order over options, not a difference order.[4]

2 Beyond Bernoulli: dropping intrinsic risk neutrality

Daniel Bernoulli (1738) famously postulated that one should maximise one's expected welfare, not one's expected monetary wealth. Like us, he used welfare as a primitive ob-

[3] in the sense that $(x, y)R(x', y') \Leftrightarrow W(x) - W(y) \geq W(x') - W(y')$ for all alternatives x, y, x', y'.

[4] Someone's choices still reveal an *incomplete fragment* of her difference order (Baccelli 2024). A difference order R is just one example of a richer structure that choices do not fully reveal and that measurement theorists have aimed to measure numerically. Another example is a so-called *extensive structure*, consisting of an ordinary binary relation (rather than a difference order) together with a 'concatenation operation' $\oplus$ that combines any two alternatives x, y into a new one $x \oplus y$, where $x \oplus y$ (intuitively) contains everything that exists in x and everything that exists in y. Nebel (2023, 2024b) measures welfare based on such a structure.

ject – he called it 'utility', but it corresponds to our 'welfare' or 'intrinsic utility'.[5] While he did not pursue our objective of 'observing' someone's welfare through her preferences, his welfare-based choice principle ties preferences so closely to welfare that welfare becomes *partially observable*. Specifically, on his approach, welfare must effectively coincide with VNM utility; and so welfare is observable to the extent that the VNM utility function is unique, i.e., up to two parameters.

Bernoulli however makes a heavy implicit assumption: the agent is neutral to intrinsic risk, i.e., to risk in welfare. We will drop this assumption, by allowing *any* intrinsic risk attitude, as long as this attitude is constant. As we will see, welfare then stays partially observable, but now a third open parameter appears, namely the intrinsic risk attitude. The question of how to pin down the three parameters of welfare will be postponed to Section 4.

We fix a set X of *situations*, in which the welfare of a given person is to be measured. While we could work without any assumptions on X (as shown in the appendix), the main text takes situations to be real numbers or more generally vectors of real numbers. Typical real-valued situations are wealth levels, health levels, or consumption index levels. Typical vector-valued situations are consumption bundles, wealth-health-education triples, or vectors of functionings (Sen 1985). Technically, the main text lets X be a non-empty open connected subset of $\mathbb{R}^k$ for some $k \geq 1$, e.g., $\mathbb{R}^k$, $(0,\infty)^k$ or $(0,1)^k$. Readers can focus on the base-line case $k = 1$, so that X is a non-empty open interval, e.g., $\mathbb{R}$, $(0,\infty)$ or $(0,1)$.

A *welfare function* or *intrinsic utility function* is a function $W : X \to \mathbb{R}$, where $W(x)$ represents the person's welfare in or from x. Welfare is not directly observed. Instead we observe ordinal comparisons between risky prospects, representing actions or policies. Let $\mathcal{P}$ be the set of *prospects*, i.e., lotteries over X with finite support. The observable primitive is a binary relation $\succeq$ on $\mathcal{P}$, called a *prospect order*, where '$x \succeq y$' means that x is observably at least as good as y for the person. The source of observation might be choice behaviour, self-assessments, third-party assessments, or perhaps neurophysiological data.

While we interpret $\succeq$ and W mainly as capturing *welfare* comparisons and *welfare* levels, one can alternatively interpret $\succeq$ and W as capturing *(weak) preference* comparisons and *preference* intensity/strength. We will take the liberty of moving back and forth between the welfare interpretation and the preference interpretation. While our terms 'prospect order' for $\succeq$ and 'intrinsic utility function' for W are neutral between both interpretations, our frequent term 'welfare function' for W suggests the welfare interpretation. Meanwhile we will frequently use preference jargon when discussing the order $\succeq$, following the economic tradition of working with 'preference orders' rather than 'welfare orders'. Our fluctuation between welfare talk and preference talk deserves an apology, because the distinction between the two is fundamental (cf. Hausman 2012). In practice, when collecting the data that give rise to $\succeq$, it is crucial to be aware whether these data reveal welfare comparisons or preferences. For instance, *choice* data normally

[5] Bernoulli used the Latin term 'emolumentum', which became translated into 'utility'. In fact, this translation is questionable, as a reviewer kindly pointed out (see Broome 1991a).

reveal a *preference* relation $\succeq$ – they reveal welfare comparisons only if we can assume that the agent pursues her own welfare in her choices, rather than pursuing altruistic objectives, or 'temptations', etc. If the data instead consist in comparative welfare assessments made by the agent or a third party, then we can more safely give $\succeq$ a welfare interpretation.

Let $\succ$ and $\sim$ denote the strict relation and indifference relation corresponding to $\succeq$; '$x \succ y$' and '$x \sim y$' mean that x is observably better than resp. as good as y for the person.[6] We identify any situation $x \in X$ with the riskless prospect that yields x for sure. So, $X \subseteq \mathcal{P}$.

How can we learn W from $\succeq$? The common move would be to identify welfare with VNM utility. A *VNM utility representation* of $\succeq$ is a function $U : X \to \mathbb{R}$ such that prospects are ranked by expected utility, i.e., for all prospects $p, q \in \mathcal{P}$, $p \succeq q \Leftrightarrow \mathbb{E}_p(U) \geq \mathbb{E}_q(U)$.

This common identification of welfare with VNM utility relies implicitly on adopting the following hypothesis, which can be ascribed to Daniel Bernoulli (1738):

Intrinsic Risk Neutrality: Any prospect is as good as getting for sure a welfare identical to the prospect's expected welfare. Formally, for any prospect $p \in \mathcal{P}$ and (riskless) outcome $x \in X$ such that $\mathbb{E}_p(W) = W(x)$, we have $p \sim x$.

This Bernoullian assumption is slightly weaker than what is often called *Bernoulli's Hypothesis*: prospects can be compared by their expected welfare, i.e., W is effectively a VNM utility function for $\succeq$.[7] In fact, also Intrinsic Risk Neutrality forces us to identify welfare with VNM utility, under a well-behavedness assumption on the welfare function. We call a function $W : X \to \mathbb{R}$ *regular* if it is smooth[8] with a nowhere zero derivative W', and *well-behaved* w.r.t. $\succeq$ if it is moreover compatible with riskless comparisons, i.e., satisfies $W(x) \geq W(y) \Leftrightarrow x \succeq y$ for all (riskless) situations $x, y \in X$.

Proposition 1 *Given a prospect order $\succeq$, a well-behaved welfare function W satisfies Intrinsic Risk Neutrality if and only if it is a VNM utility function, i.e., takes the form*

$$W = U$$

for some VNM representation U of $\succeq$.[9]

Bernoulli's Intrinsic Risk Neutrality was a significant progress at the time: it overcame the naive idea of neutrality to risk *in outcomes*, replacing it with neutrality to risk *in*

[6] For all $p, q \in \mathcal{P}$, $p \succ q$ if and only if $p \succeq q$ and not $q \succeq p$, and $p \sim q$ if and only if $p \succeq q$ and $q \succeq p$.

[7] Formally: for any prospects $p, q \in \mathcal{P}$, $p \succeq q \Leftrightarrow \mathbb{E}_p(W) \geq \mathbb{E}_q(W)$.

[8] 'Smooth' means that W is differentiable arbitrarily many often. In the multi-dimensional case $X \subseteq \mathbb{R}^k$ with $k \geq 2$, W' is of course a vector $(\frac{d}{dx_1}W, ..., \frac{d}{dx_k}W)$, and W' is 'nowhere zero' if at no $x \in X$ it is the zero vector $(0, ..., 0)$.

[9] In other words: given a prospect order $\succeq$, if a welfare function W is well-behaved, then Intrinsic Risk Neutrality is equivalent to Bernoulli's Hypothesis.

welfare.[10] What is worrying is that Bernoulli still relied on a neutral attitude to risk (now understood as *intrinsic* risk). He fixed only one of two problems: he rightly made welfare the 'currency' of risk, but he retained a neutral attitude to risk. We will thus replace his Intrinsic Risk Neutrality with a more general hypothesis, which still makes welfare the currency of risk, but allows any attitude to risk – neutrality, aversion or proneness – as long as this attitude is stable, i.e., independent of the situation or welfare level. Here is our condition, where we call a welfare $w \in \mathbb{R}$ *equivalent* to a prospect $p \in \mathcal{P}$ if it is achieved in a situation that is as good as p, i.e., if $w = W(x)$ for a situation $x \sim p$.[11]

Constant Intrinsic Risk Attitude (CIRA): If a prospect is modified by increasing each possible resulting *welfare* by the same amount, then the equivalent *welfare* also increases by this amount. Formally, for all $\Delta > 0$ and all prospects $p \in \mathcal{P}$ with an equivalent welfare $w_p \in \mathbb{R}$, if another prospect $q \in \mathcal{P}$ with an equivalent welfare $w_q \in \mathbb{R}$ satisfies $q(W = w + \Delta) = p(W = w)$ for all $w \in \mathbb{R}$, then $w_q = w_p + \Delta$.[12]

For instance, if getting welfare 1 or 2 with equal probability is as good as getting welfare 1.4 for sure, then getting welfare 2 or 3 with equal probability is as good as getting welfare 2.4 for sure.

Fact 1: *CIRA generalises Bernoulli's Intrinsic Risk Neutrality.*

CIRA requires a 'coherent' or 'stable' attitude to intrinsic risk. For instance, if at low current welfare the agent is indifferent towards a 50:50 gamble that lets her gain 2 *welfare* units or lose 1 *welfare* unit, then she stays indifferent towards this gamble at higher initial welfare.

Just as Bernoulli's Intrinsic Risk Neutrality improves on an outcome-based concept of risk neutrality, CIRA also improves on an outcome-based concept, namely the concept of Constant Absolute Risk Aversion or 'CARA'. The improvement consists again in making welfare the currency of risk. The advantage of CIRA over CARA will be discussed in a moment. Much later, we will present a more systematic argument for CIRA, by deriving CIRA from two principles, namely that the preferences are dynamically consistent and difference-based (Section 7).

By being more general than Bernoulli's Intrinsic Risk Neutrality, CIRA allows for a wider class of welfare functions. While Intrinsic Risk Neutrality forces one to measure welfare by VNM utility (Proposition 1), CIRA implies the following welfare measure. We

[10] *Outcome risk neutrality* (for $X \subseteq \mathbb{R}$): Any prospect is as good as getting for sure the *outcome* identical to the prospect's expected *outcome*. Formally, for any prospect $p \in \mathcal{P}$ and (riskless) outcome $y \in X$ such that $\sum_{x \in X} p(x)x = y$, we have $p \sim y$.

[11] Except in degenerate cases, each prospect has a unique equivalent welfare. This is for instance guaranteed if W is well-behaved and $\succeq$ is regular (as defined in a moment).

[12] There is an equivalent statement of CIRA (assuming prospects have unique equivalent welfares): If a prospect is modified by a fixed increase in welfare, then the intrinsic risk premium is unchanged. By the *intrinsic risk premium* of a prospect $p \in \mathcal{P}$ we mean the gap $\mathbb{E}_p(W) - w_p$ between p's equivalent welfare w_p and p's expected welfare $\mathbb{E}_p(W)$.

will assume that the prospect order $\succeq$ is *regular*, i.e., has a regular VNM representation U.

Proposition 2 *Given a regular prospect order $\succeq$, a well-behaved welfare function W satisfies CIRA if and only if it takes the form*

$$W = \log(\rho U + 1)/\rho$$

for some VNM representation U of $\succeq$ and some 'intrinsic risk attitude' $\rho \in \mathbb{R}$ with $\rho U + 1 > 0$.

If $\rho = 0$, then '$\log(\rho U + 1)/\rho$' stands for U $(= \lim_{\rho \to 0} \log(\rho U + 1)/\rho)$. So, for $\rho = 0$ we obtain Intrinsic $W = U$, which is the special case of Intrinsic Risk Neutrality, treated in Proposition 1. In this sense, Proposition 2 generalises Proposition 1.

In general:

- if $\rho = 0$, the agent is *intrinsic risk neutral*, as VNM utility equals welfare,
- if $\rho > 0$, the agent is *intrinsic risk prone*, as VNM utility is convex in welfare (since welfare is concave in VNM utility)
- if $\rho < 0$, the agent is *intrinsic risk averse*, as VNM utility is concave in welfare (since welfare is convex in VNM utility).

Although the exact value of ρ is usually empirically underdetermined based on assuming CIRA, the sign of ρ – and hence the *qualitative* intrinsic risk attitude – is observable. More precisely, based on assuming CIRA, the agent is observably

- intrinsic risk neutral, i.e., $\rho = 0$, if $\sup U = \infty$ and $\inf U = -\infty$ (reason: $\rho U + 1 > 0$),
- weakly intrinsic risk prone, i.e., $\rho \geq 0$, if $\sup U = \infty$ and $\inf U \neq -\infty$ (reason: $\rho U + 1 > 0$),
- weakly intrinsic risk averse, i.e., $\rho \leq 0$, if $\sup U \neq \infty$ and $\inf U = -\infty$ (reason: $\rho U + 1 > 0$).

We note an equivalent restatement of Proposition 2:

Proposition 2 (restated) *Given a regular prospect order $\succeq$, a well-behaved welfare function W satisfies CIRA if and only if*

- *either $W = \sigma \log V$ for some VNM representation V of $\succeq$ with $V > 0$ and some $\sigma > 0$ (case of intrinsic risk proneness)*
- *or $W = -\sigma \log(-V)$ for some VNM representation V of $\succeq$ with $V < 0$ and some $\sigma > 0$ (case of intrinsic risk aversion)*
- *or $W = V$ for some VNM representation V of $\succeq$ with full range $\mathbb{R}$ (case of intrinsic risk neutrality).*

This restatement of Proposition 2 uses a different parametrisation of the welfare function, by rescaling the VNM function U into V and replacing the risk-attitudinal parameter ρ with σ:

- If the agent is intrinsic risk prone, i.e., $\rho > 0$, then $\sigma = 1/\rho$ and $V = \rho U + 1$.
- If the agent is intrinsic risk averse, i.e., $\rho < 0$, then $\sigma = -1/\rho$ and $V = -(\rho U + 1)$.
- If the agent is intrinsic risk neutral, i.e., $\rho = 0$, then $V = U$.

The alternative parametrisation highlights the three different functional forms of welfare depending on the intrinsic risk attitude, i.e., on the sign of ρ. We will continue to work with the initial parametrisation $W = \log(\rho W + 1)/\rho$, because it is unified and explicitly features the intrinsic risk attitude ρ.

Our discussion of risk attitudes is an example of our (philosophically unfortunate) fluctuation between the welfare interpretation and the preference interpretation of our model. On the one hand, we invoke risk in resulting welfare, which suggests that W captures welfare. On the other hand, we invoke terms like 'risk attitude' and 'risk aversion', which suggests that $\succeq$ captures attitudes of the agent, hence preferences rather than welfare. All this can however be made consistent.[13]

Excursion: Why CIRA is more plausible than standard CARA – and can explain empirical violations of CARA

CIRA contrasts with the following well-known condition, which operates at the level of outcomes rather than welfare. As usual, an outcome $x \in \mathbb{R}$ is called *equivalent* to a prospect $p \in \mathcal{P}$ (or a *certainty equivalent* of p) if x is as good as p, i.e., $x \sim p$.

Constant Absolute Risk Aversion (CARA), defined for $X \subseteq \mathbb{R}$: If a prospect is modified by increasing each possible resulting *outcome* by the same amount, then the equivalent *outcome* also increases by this amount. Formally, for all $\Delta > 0$ and all prospects $p \in \mathcal{P}$ with an equivalent outcome $x_p \in X$, if another prospect $q \in \mathcal{P}$ with an equivalent outcome $x_q \in X$ satisfies $q(x + \Delta) = p(x)$ for all $x \in \mathbb{R}$, then $x_q = x_p + \Delta$.[14]

[13] Proponents of a preference interpretation should replace 'risk in welfare' (how well off will she be?) with 'risk in preference satisfaction' (how much will her preferences be satisfied?). The currency of risk is then preference satisfaction. Proponents of the welfare interpretation should do something else: they should read terms like 'risk attitude' metaphorically, as standing for the (positive, negative, or neutral) effect of riskiness on welfare. Incidentally, this purely metaphorical risk attitude might be causally affected by the agent's real risk attitude – under natural theories of welfare, such as desire-satisfaction theories, happiness theories, or constitutively plural theories. Indeed, someone who is risk-averse in the ordinary preferential sense will presumably suffer a *welfare* loss from risk, hence be risk-averse in our metaphorical sense too.

[14] To make '$p(x)$' and '$q(x + \Delta)$' well-defined even if x resp. $x + \Delta$ fall outside X, identify any lottery over X ($\subseteq \mathbb{R}$) with its extension to $\mathbb{R}$, which is zero within $\mathbb{R} \backslash X$. There is an equivalent statement of CARA (assuming prospects have unique certainty equivalents): If a prospect is modified by a fixed increase in outcomes, then the risk premium is unchanged. The *risk premium* of a prospect $p \in \mathcal{P}$ is the gap $\overline{p} - x_p$ between p's certainty equivalent x_p and p's expectation $\overline{p} = \sum_{x \in X} p(x)x$.

For instance, assuming outcomes are wealth levels, if owning \$0 or \$10,000 with equal probability is as good as owning \$2,000 for sure, then owning \$100,000 or \$110,000 with equal probability must be as good as owning \$102,000 for sure. CARA is implausible, as is confirmed by empirical violations (Chiappori and Paiella 2011). Why? If a risky wealth prospect is translated upwards, then it moves into a region of higher wealth and thus lower marginal welfare, assuming diminishing marginal welfare. So the new prospect contains less *intrinsic* risk, i.e., less risk in the subjectively relevant sense of welfare. This leads to a smaller risk premium, assuming intrinsic risk aversion – a violation of CARA. For instance, a 50:50 lottery between wealth \$0 and wealth \$10,000 contains huge intrinsic risk: $W(0) \ll W(10,000)$. But the translated 50:50 lottery between wealth \$100,000 and \$110,000 contains little intrinsic risk: $W(100,000) \approx W(110,000)$. So, the first lottery should be equivalent to a wealth close to the low outcome \$0, and the second to a wealth close to the average outcome $\$105,000$, violating CARA.

CIRA is not subject to this sort of objection, since it accounts for decreasing marginal welfare by considering 'welfare shifts' rather than 'outcome shifts'. In fact, CIRA can offer a systematic explanation for the well-documented empirical violation of CARA: *CIRA rules out CARA*, except if welfare is of a very special form. To state this result, we call a real function on a subset of $\mathbb{R}$ *linear* if it is given by $x \mapsto ax + b$ for some $a, b \in \mathbb{R}$, *exponential* if it is given by $x \mapsto ae^{bx} + c$ for some $a, b, c \in \mathbb{R}$ with $a, b \neq 0$, and *logarithmic* if it is given by $x \mapsto a \log(bx + c)$ for some $a, b, c \in \mathbb{R}$ with $a, b \neq 0$.

Fact 2 *(informal statement[15]): CIRA rules out CARA provided the welfare function is neither linear, nor exponential, nor logarithmic, nor a logarithmic function of an exponential function.*

The natural Bernoullian response to empirical violations of CARA would be to make welfare the currency of risk. This is exactly what we do by replacing CARA with CIRA. Modern choice theorists instead replace CARA with some other *outcome*-level hypothesis, such as 'hyperbolic absolute risk aversion' (HARA). This different response reflects different objectives: while most choice theorists aim to represent or predict choice, we aim to make welfare indirectly observable. The former objective precludes using conditions like CIRA that refer to unobservables, according to orthodox choice theory. Our objective instead *requires* using conditions that relate the relevant unobservable (W) to the observable ($\succeq$), according to the established scientific methodology for identifying unobservables.[16]

[15] The exact characterisation is stated in the appendix as Proposition 8.

[16] Applied economists, statisticians, psychologists, physicists and other empirical scientists all routinely rely on this approach: they all make inferences about relevant unobservables from observables via theoretic hypotheses linking the two. Of course, each field has its own type of observables, unobservables and hypotheses.

3 Explaining standard utility and risk attitude by intrinsic utility and risk attitude

Before continuing with the measurement of welfare, let us explore the structure of classic VNM utility and classic Arrow-Pratt risk attitude, by decomposing both objects into their two fundamental determinants, welfare and intrinsic risk attitude. Our decompositions will show that standard utility and risk attitude are hybrid constructs, resulting from an interplay of two distinct ingredients. At this stage, the decomposition will not be fully observable, because welfare still has some open parameters in Proposition 2. Unique identification will be achieved later.

Arrow-Pratt's classic theory measures risk attitudes as follows, and needs to assume one-dimensional outcomes:

Definition 1 *The **classical (or Arrow-Pratt) risk attitude** of a regular prospect order $\succeq$, for $X \subseteq \mathbb{R}$, is the (well-defined[17]) function $\rho_{AP} = \rho_{\succeq,AP} = \frac{U''}{U'}$, where U is any VNM representation of $\succeq$. If constant, ρ_{AP} is identified with its single value.*

We measure the intrinsic risk attitude analogously, by simply replacing outcomes with welfare levels, no longer having to assume one-dimensional outcomes:

Definition 2 *The **intrinsic risk attitude** of a regular prospect order $\succeq$ w.r.t. a well-behaved welfare function W is the (well-defined[18]) function $\rho_W = \rho_{\succeq,W} = \frac{d^2U/dW^2}{dU/dW}$, where U is any VNM representation of $\succeq$. If constant, ρ_W is identified with its single value.*

Here is a well-known fact about Arrow-Pratt risk attitudes, and its (similarly provable) analogue for intrinsic risk attitudes:

Remark 1 *$\succeq$ satisfies CARA if and only if its classical risk attitude ρ_{AP} is constant.*

Remark 2 *$\succeq$ satisfies CIRA w.r.t. W if and only if its intrinsic risk attitude ρ_W is constant.*

By simple algebra, the formula '$W = \log(\rho U + 1)/\rho$' in Proposition 2 implies that $U = (e^{\rho W} - 1)/\rho$, and hence that $\rho_W = \rho$. So, Proposition 2 has two corollaries. First, we can replace ρ with ρ_W in the formula '$W = \log(\rho U + 1)/\rho$':

Corollary 1 *In the welfare function $W = \log(\rho U + 1)/\rho$ in Proposition 2, ρ equals the (constant) intrinsic risk attitude ρ_W.*

[17] As $\succeq$ is regular, $\frac{U''}{U'}$ is well-defined, i.e., U exists and is twice differentiable with $U' \neq 0$.

[18] Well-definedness of $\frac{d^2U/dW^2}{dU/dW}$ requires that U be twice differentiable in W with nowhere zero first derivative in W, more precisely that U be writeable as $\phi(W)$ for a (unique) function $\phi : Rg(W) \to \mathbb{R}$ that is twice differentiable with nowhere zero ϕ' (in which case dU/dW stands for $\phi'(W)$ and d^2U/dW^2 stands for $\phi''(W)$). Well-definedness follows from the regularity of $\succeq$ and W (in fact, ϕ' is everywhere positive, as can be seen via Lemma 1).

Second, VNM utility can be explained by two determinants:

Corollary 2 *A regular prospect order $\succeq$ has a VNM utility representation U determined by welfare and the intrinsic risk attitude via*

$$U = (e^{\rho_W W} - 1)/\rho_W,$$

given any well-behaved welfare function W satisfying CIRA.

As usual, if $\rho_W = 0$, i.e., if $\succeq$ is intrinsic risk neutral, then '$(e^{\rho_W W} - 1)/\rho_W$' stands for W $(= \lim_{\rho \to 0} \log(e^{\rho W} - 1)/\rho)$.

Not only VNM utility, but also the classical risk attitude ρ_{AP} can be explained in terms of an interplay between welfare and the intrinsic risk attitude, this time without having to postulate CIRA:

Proposition 3 *The classical risk attitude of a regular prospect order $\succeq$, with $X \subseteq \mathbb{R}$, is determined by welfare and the intrinsic risk attitude via*

$$\rho_{AP} = \frac{W''}{W'} + W' \rho_W,$$

given any well-behaved welfare function W.

Thus the classical risk attitude $\rho_{AP} = \frac{U''}{U'}$, i.e., the growth rate of marginal utility, is the sum of two things:

- a 'welfare component' $\frac{W''}{W'}$, i.e., the growth rate of marginal welfare,
- a 'risk component' $W' \rho_W$, i.e., the intrinsic risk attitude weighted by marginal welfare. The weighting by marginal welfare reflects the fact that risk in welfare matters to the extent that welfare varies, i.e., to the extent W'.

Since Proposition 3 does not require CIRA, the intrinsic risk attitude ρ_W can be non-constant, and VNM utility U need not equal $(e^{\rho_W W} - 1)/\rho_W$. Still U must obey a differential equation: $\frac{U''}{U'} = \frac{W''}{W'} + W' \rho_W$. So VNM utility stays determined by welfare (W) and intrinsic risk attitude (ρ_W). Thus the central conceptual point – that classical utility has two distinct determinants – does not hinge on CIRA.

4 Fully measurable welfare

So far, the welfare function W is only partially revealed by the observable $\succeq$. More precisely, the welfare function W in Proposition 2 has three open parameters: the intrinsic risk attitude ρ and the two open parameters implicit in the choice of VNM representation U. Surprisingly, full uniqueness of W can be achieved by adding two simple hypotheses about welfare, namely a range condition and a normalisation condition. We begin with the range condition:

Full-range: There exist arbitrarily good or bad situations. That is, for all $w \in \mathbb{R}$ there is a situation $x \in X$ in which welfare is $W(x) = w$.

Full-range is a richness assumption on the set of situations considered: this set should include situations of arbitrary quality, be these situations realistic or merely theoretic. Note that VNM utility could still be bounded below or above.

Implicitly, Full-range is also a condition on the scale on which welfare is measured: that scale should include all real numbers as meaningful welfare levels. We return to measurement-theoretic issues later, when we will generalise Full-range to allow for other measurement scales.

Surprisingly, by adding the condition Full-range we settle the value of the risk attitude ρ in the formula '$W = \log(\rho U + 1)/\rho$'. The next proposition shows this. It will assume that the prospect order $\succeq$ is *broad-ranging*. This means that for any situations there exist much better or much worse situations, more precisely: for any $x, y \in X$ with $x \succeq y$ there exists a $z \in X$ such that $z_{\frac{1}{2}}y_{\frac{1}{2}} \succ x$ or $y \succ z_{\frac{1}{2}}x_{\frac{1}{2}}$. Here, '$z_{\frac{1}{2}}y_{\frac{1}{2}} \succ x$ or $y \succ z_{\frac{1}{2}}x_{\frac{1}{2}}$' means that z is *either* so good that its 50:50 mixture with y beats x *or* so bad that its 50:50 mixture with x loses to y. Broad-rangingness holds in most standard models of preferences under risk, including all HARA models.[19]

Remark: *Any VNM representation U of a broad-ranging prospect order $\succeq$ is unbounded, i.e., satisfies $\sup U = \infty$ or $\inf U = -\infty$.*

Proposition 4 *Given a regular and broad-ranging prospect order $\succeq$, a well-behaved welfare function W satisfies CIRA and Full-range if and only if it takes the form*

$$W = \log(\rho U + 1)/\rho$$

for some (unbounded) VNM representation U of $\succeq$ with $0 \in Rg(U)$ and the 'intrinsic risk attitude'

$$\rho = \begin{cases} \frac{-1}{\sup U} \ (< 0) & \text{if } \sup U \neq \infty & \text{(intrinsic risk aversion)} \\ \frac{-1}{\inf U} \ (> 0) & \text{if } \inf U \neq -\infty & \text{(intrinsic risk proneness)} \\ 0 & \text{if } \sup U = \infty \text{ and } \inf U = -\infty & \text{(intrinsic risk neutrality).} \end{cases}$$

By now, only the choice of U is open. To settle U, we will normalise welfare. A popular idea in the theory of fair allocation is that individual welfare is set to a fixed 'objective' level in particular situations, such as situations on the poverty line, situations of perfect health, or social situations in which all goods are distributed equally (Fleurbaey and Maniquet 2011). We will do something similar. Consider a fixed reference situation $\overline{x} \in X$, representing for instance a 'poverty point'. A function from X to $\mathbb{R}$ (such as W) is *normalised* if at the reference point $\overline{x}$ it takes the value 0 and has a derivative of size

[19] For all distinct $x, y \in X$ and all $t \in [0, 1]$, $x_t y_{1-t}$ denotes the prospect $p \in \mathcal{P}$ such that $p(x) = t$ and $p(y) = 1 - t$.

1. The derivative of W, or *marginal welfare*, captures the effect of small external changes on welfare.[20]

Normalisation: The welfare function W is normalised.

This condition has a different status from CIRA and Full-range, by being of a more conventional nature. One could work without Normalisation. Then welfare is non-unique, but still cardinal, as will be shown later – just as unnormalised VNM utility is non-unique, but still cardinal.

Normalisation requires measuring welfare on a scale that sets welfare to 0 and marginal welfare size to 1 at the reference point. A measurement scale is a convention that fixes the meaning of numbers, by setting a correspondence between numbers and real-world welfare states. One can always scale welfare such that Normalisation holds: every well-behaved welfare function satisfying CIRA and Full-range can be transformed into one that additionally satisfies Normalisation, by applying an increasing affine transformation. The fact that rescaling a welfare function changes welfare levels and welfare differences does not make welfare levels and differences meaningless. Rather it makes the meaning of levels and differences scale-relative: statements such as 'welfare is 2' and 'welfare rises by 3' can still have meanings, which are fixed by the chosen scale.

Still, Normalisation is far from innocent, for two reasons. For one, Normalisation becomes questionable when one engages in interpersonal comparisons of welfare levels and/or differences, because meanings cannot be fixed in more than one way. Normalisation effectively takes people to be welfare-wise *similar at the reference point*. Here is the second problem. A welfare analyst normally wants to *know* the convention or scale on which she measures the agent's welfare. She wants to know which real-world welfare states are being denoted by numbers such as 0 or 7. But if she does not know the agent's real welfare state at and near the reference point, say because she only knows the welfare comparisons given by $\succeq$, then she does not know the new scale or convention that she is enforcing by accepting Normalisation. This is a serious problem – conventions are arbitrary, but they'd better be known. Perhaps surprisingly, Normalisation is nonetheless sometimes defensible, as argued in Section 7.

We now state our central theorem, whereby our three hypotheses lead to a unique, and thus revealed, welfare function, obtained by choosing U and ρ in particular ways in the formula '$W = \log(\rho U + 1)/\rho$':

Theorem 1 *Given a regular and broad-ranging prospect order $\succeq$, there exists a unique well-behaved welfare function W satisfying CIRA, Full-range, and Normalisation, namely the function*

$$W = \log(\rho U + 1)/\rho$$

[20] In the basic case $X \subseteq \mathbb{R}$, the size W' is the absolute value $|W'|$, which normally equals W' as $W' > 0$, i.e., as 'more is better'. In the general case $X \subseteq \mathbb{R}^k$ $(k \geq 1)$, the size of $W' = \left(\frac{dW}{dx_1}, \ldots, \frac{dW}{dx_k}\right)$ is the length $\|W'\|$.

based on the normalised (unbounded) VNM representation U of $\succeq$ and the 'intrinsic risk attitude'

$$\rho = \begin{cases} \frac{-1}{\sup U} \ (<0) & \text{if } \sup U \neq \infty & \text{(intrinsic risk aversion)} \\ \frac{-1}{\inf U} \ (>0) & \text{if } \inf U \neq -\infty & \text{(intrinsic risk proneness)} \\ 0 & \text{if } \sup U = \infty \text{ and } \inf U = -\infty & \text{(intrinsic risk neutrality).} \end{cases}$$

As usual, if $\rho = 0$, then '$\log(\rho U + 1)/\rho$' stands for U $(= \lim_{\rho \to 0} \log(\rho U + 1)/\rho)$. An alternative statement of the result draws on the notion of *revealed* welfare and *revealed* intrinsic risk attitude.

Definition 3 *A prospect order $\succeq$*

- ***reveals the welfare function*** *W if W is the only well-behaved welfare function satisfying CIRA, Full-range and Normalisation, in which case W is denoted $W_{\succeq}$,*
- ***reveals the intrinsic risk attitude*** *ρ if $\succeq$ reveals a welfare function $W_{\succeq}$ and $\rho = \rho_{W_{\succeq}}$, in which case ρ is denoted $\rho_{\succeq}$.*

Theorem 1 *(restated) Every regular and broad-ranging prospect order $\succeq$ reveals a welfare function and a constant intrinsic risk attitude, given by*

$$W_{\succeq} = \log(\rho_{\succeq} U + 1)/\rho_{\succeq}$$

and

$$\rho_{\succeq} = \begin{cases} \frac{-1}{\sup U} \ (<0) & \text{if } \sup U \neq \infty & \text{(intrinsic risk aversion)} \\ \frac{-1}{\inf U} \ (>0) & \text{if } \inf U \neq -\infty & \text{(intrinsic risk proneness)} \\ 0 & \text{if } \sup U = \infty \text{ and } \inf U = -\infty & \text{(intrinsic risk neutrality)} \end{cases}$$

where U is the normalised (unbounded) VNM representation of $\succeq$.

Our definition of 'revealed' welfare (and risk attitude) contains a bias, by adopting particular hypotheses, namely CIRA, Full-range, and Normalisation. While we do regard these hypotheses as particularly salient, other sets of hypotheses are imaginable. One could for instance drop Normalisation. A more explicit terminology would be to talk of the welfare that is 'revealed *w.r.t. such-and-such hypotheses*'. By varying the hypotheses, we would then obtain different hypothesis-dependent notions of (full or partial) revelation or measurability of welfare. Section 6 discusses all this.

Note a striking disanalogy between the classical and intrinsic risk attitudes, ρ_{AP} and $\rho_{\succeq}$. While both can be derived from a VNM representation U, ρ_{AP} $(= \frac{U''}{U'})$ is derived *locally*, from the curvature of U, whereas $\rho_{\succeq}$ is derived *globally*, from the range of (normalised) U.

Empirical application

To apply Theorem 1 in practice, one should first choose a (normalised) VNM function U that fits the data $\succeq$ as well as possible, and then derive $W_{\succeq}$ and $\rho_{\succeq}$ via Theorem 1. For instance, U could be chosen from the class of HARA utility functions. These (unbounded) utility functions are empirically well-confirmed. This opens the door for a social welfare analysis based on more convincing individual welfare measures than VNM utility functions.

For example, let $X = (0, \infty)$, and assume the data $\succeq$ support decreasing absolute risk aversion, with a constant *relative* risk aversion of $\eta \geq 0$. This 'CRRA' model – a special case of HARA – has been widely confirmed, although the value of η is context-dependent. The agent's normalised utility function is then of the well-known CRRA type:

$$U(x) = \frac{\overline{x}}{1-\eta}\left(\left(\frac{x}{\overline{x}}\right)^{1-\eta} - 1\right) \text{ for all } x \in X.$$

If $\eta = 1$, this formula is interpreted as $U(x) = \overline{x} \log \frac{x}{\overline{x}}$ $(= \lim_{\eta \to 1} \frac{\overline{x}}{1-\eta}\left(\left(\frac{x}{\overline{x}}\right)^{1-\eta} - 1\right))$. Applying the formulas in Theorem 1, one finds that the agent has a revealed intrinsic risk attitude given by[21]

$$\rho_{\succeq} = \frac{1-\eta}{\overline{x}},$$

and a revealed welfare given by[22]

$$W_{\succeq}(x) = \overline{x} \log \frac{x}{\overline{x}} \text{ for all } x \in X.$$

So, the well-known CRRA model implies a special logarithmic welfare.[23] Other utility models than the CRRA model – including other HARA models – would lead to other concrete formulas for welfare and intrinsic risk attitude via Theorem 1.

5 Measuring social welfare

We now turn to *social* welfare, as a guide to policy choices. Let us rely on the welfarist idea that social welfare is a function of people's individual welfare. This section will contrast two approaches to (welfarist) social welfare analysis: the standard approach where individual welfare is measured by VNM utility, and a new approach based on our individual welfare measure.

The section interprets X as containing *social* situations, faced by individuals $i = 1, ..., n$ $(n \geq 2)$. Each person i holds an (observed) prospect order $\succeq_i$, representing her

[21] Check this by distinguishing between the cases $\eta > 1$ (where $\sup U < \infty$), $\eta < 1$ (where $\inf U > -\infty$) and $\eta = 1$ (where $\sup U = \infty$ and $\inf U = -\infty$).

[22] Since $W(x) = \frac{\log(\rho U(x)+1)}{\rho} = \frac{\log\left(\rho \frac{1}{\rho}\left(\left(\frac{x}{\overline{x}}\right)^{\rho\overline{x}} - 1\right)+1\right)}{\rho} = \frac{\log\left(\left(\frac{x}{\overline{x}}\right)^{\rho\overline{x}}\right)}{\rho} = \overline{x} \log \frac{x}{\overline{x}}$.

[23] Interestingly, this welfare function is independent of the relative risk aversion η. Thus the debate about the right value of η does not affect revealed welfare – it only affects VNM utility and choice. Welfare is subject to less measurement uncertainty than VNM utility in this model, despite being revealed more indirectly.

ranking under risk. We take each $\succeq_i$ to be regular and broad-ranging, so that each $\succeq_i$ reveals a welfare function, given by $W_i = \log(\rho_i U_i + 1)/\rho_i$, where ρ_i is i's revealed intrinsic risk attitude and U_i is i's normalised VNM utility function. An important application is a *vector model*: here, social situations are vectors of individual situations, i.e., $X = X_1 \times \cdots \times X_n$, where X_i contains person i's possible individual situations (such as wealth levels or consumption bundles), and where $\succeq_i$ and W_i depend only on i's situation, hence are essentially an order or function on X_i rather than X.[24] In typical vector models, all X_i's coincide.

How well-off is society? We will begin with the social evaluation of riskless situations (Section 5.1), followed by the social evaluation of risky prospects (Section 5.2), and a methodological discussion (Section 5.3).

5.1 Social welfare under certainty

Let $W : X \to \mathbb{R}$ be a *social* welfare function. How should it be defined? We will limit ourselves to two approaches, utilitarianism and prioritarianism (see Bossert and Weymark 2004 and Adler 2019 for overviews). These two approaches can each be defined in two ways, depending on how personal welfare is measured:

W-Utilitarianism: $W = \sum_i W_i$ for the welfare functions W_i revealed by the orders $\succeq_i$.

VNM-Utilitarianism: $W = \sum_i U_i$ for some VNM representations U_i of the orders $\succeq_i$.

W-Prioritarianism: $W = \sum_i \pi(W_i)$ for the welfare functions W_i revealed by the orders $\succeq_i$ and some strictly concave transformation $\pi : \mathbb{R} \to \mathbb{R}$.

VNM-Prioritarianism: $W = \sum_i \pi(U_i)$ for some VNM representations U_i of the orders $\succeq_i$ and some strictly concave transformation $\pi : \mathbb{R} \to \mathbb{R}$.

The two VNM-based social welfare theories do not uniquely specify social welfare comparisons, leaving open the choice of the VNM functions U_i. This non-uniqueness is not the real disadvantage compared to W-based social welfare theories. VNM-based theories could achieve uniqueness too, by adding the same Normalisation condition that is also built into W-based accounts, thereby forcing each U_i to be the normalised VNM function. Let us call the resulting theories *Normalised VNM-Utilitarianism* and *Normalised VNM-Prioritarianism*. If one instead rejects the Normalisation condition, then one should remove it from W-based accounts too, so that the W-based theories would become similarly non-unique (see Section 6 on measuring welfare without Normalisation).

The problem with VNM-Utilitarianism and VNM-Prioritarianism is that U_i does not faithfully measure i's welfare, being distorted by i's risk attitude. Intuitively, individual

[24] The reference point $\overline{x}$ is then a vector $(\overline{x}_1, \ldots, \overline{x}_n)$ of individual reference points, e.g., individual poverty points.

risk attitudes should be irrelevant to social welfare considerations *under certainty*. The two W-based accounts filter out risk attitudes by using W_i rather than U_i.

What goes wrong if a utilitarian like John Harsanyi maximises $\sum_i U_i$ rather than $\sum_i W_i$? Under the plausible assumption that people are intrinsic risk averse, each U_i is a concave transformation of W_i.[25] So, maximising total utility boils down to maximising total *concavely transformed* welfare – which is no longer utilitarianism, but a form of prioritarianism. More precisely:

Fact 3: VNM-Utilitarianism implies W-Prioritarianism, assuming that all persons are (equally[26]) intrinsic risk averse.

So, ironically, the dedicated utilitarian John Harsanyi is effectively a prioritarian.[27] And his famous 'utilitarian theorem', which supports VNM-Utilitarianism, effectively supports Prioritarianism, understood as W-Prioritarianism. More on this theorem in Section 5.2.

Prioritarianism is driven by an aversion to inequality. What is inequality aversion? Just as for risk aversion, there are two different approaches, depending on whether one construes inequality as inequality *in outcomes* (e.g., in wealth) or as inequality *in welfare*. Prioritarians care about inequality in welfare, since the transformation π is applied to welfare rather than material outcomes.[28]

It is debatable whether welfare inequality is more relevant than material inequality, and hence whether prioritarians are right to cash out inequality in terms of welfare rather than outcomes. Something should however be clear: *if* the currency of inequality is welfare *then* welfare should not be measured by VNM utility – otherwise the inequality in people's (riskless) welfare levels would depend on people's intrinsic risk attitudes.

Axiomatically speaking, prioritarians distinguish themselves from utilitarians by adopting the

Pigou-Dalton Principle: All transfers from a better-off to a worse-off are social improvements.

[25] Because the normalised VNM representation is given by $(e^{\rho_i W_i} - 1)/\rho_i$, (where ρ_i is i's intrinsic risk attitude) and this expression is concave in W_i since $\rho_i < 0$.

[26] The assumption that all individuals i have the *same* intrinsic risk attitude $\rho_i \equiv \rho$ (< 0) can be dropped if one generalises Prioritarianism by allowing the transformation π to depend on the individual i.

[27] Conversely, certain special (W-)prioritarians are effectively VNM-utilitarians. These are those prioritarians whose transformation π is tied to people's risk attitude, such that $\pi(W_i)$ coincides with a VNM utility function U_i of person i. This presupposes in particular that everyone is intrinsic risk averse (to the same degree) – otherwise π would not be strictly concave (and independent of i). Can there be any reason for a prioritarian to adopt such a transformation π, and hence to become effectively a VNM-utilitarian? Perhaps. Cibinel (2025b) indeed defends a similar version of Prioritarianism, referred to as the *risk priority view*.

[28] A non-standard 'material' prioritarian would apply a transformation π to individual *outcomes*. This presupposes a vector model, in which social outcomes/situations in X are vectors $x = (x_1, \ldots, x_n)$ of individual situations, each of which is a real number (e.g., a wealth level).

There are different versions of this principle, depending on how welfare is measured. We focus on the versions based on our welfare measure and on VNM utility:

W-Pigou-Dalton Principle: Relative to the welfare functions $W_1, \ldots, W_n$ revealed by the orders $\succeq_1, \ldots, \succeq_n$, all transfers from a better-off to a worse-off are social improvements, i.e., for all persons i, j, amounts $\Delta > 0$, and situations $x, y \in X$, if $W_i(y) = W_i(x) - \Delta$, $W_j(y) = W_j(x) + \Delta$, $W_i(y) > W_j(y)$, and $W_k(y) = W_k(x)$ for all other persons k, then $W(y) > W(x)$.

VNM-Pigou-Dalton Principle: Relative to some VNM representations $U_1, \ldots, U_n$ of the orders $\succeq_1, \ldots, \succeq_n$, all transfers from a better-off to a worse-off are social improvements, i.e., for all persons i, j, amounts $\Delta > 0$, and situations $x, y \in X$, if $U_i(y) = U_i(x) - \Delta$, $U_j(y) = U_j(x) + \Delta$, $U_i(y) > U_j(y)$, and $U_k(y) = U_k(x)$ for all other persons k, then $W(y) > W(x)$.

W-Prioritarianism satisfies the W-Pigou-Dalton Principle, whereas VNM-Prioritarianism satisfies the VNM-Pigou-Dalton Principle, and normalised VNM-Utilitarianism satisfied a version of the VNM-Pigou-Dalton Principle with normalised rather than arbitrary VNM representations $U_1, \ldots, U_n$. Proponents of *material* equality would instead adopt a 'Material Pigou-Dalton Principle'.[29]

The VNM-Pigou-Dalton Principle is questionable. Person i might only *look* better off than person j, and person j might only *seem* to gain as much welfare as person i loses, because VNM utility is notoriously distorted by (intrinsic) risk attitude. Person j might actually enjoy a better life than person i, and/or gain less real-world well-being than i loses. The VNM-Pigou-Dalton principle tends to recommend transfers from the risk-prone to the risk-averse, insofar as risk-averse people have upper-bounded and lower-unbounded VNM utility, while risk-prone people have upper-unbounded and lower-bounded VNM utility. Yet rewarding risk aversion seems ethically unjustified.

5.2 Social welfare under risk

A growing literature addresses the evaluation of social welfare under risk (see Harsanyi 1956, Fleurbaey 2010, Adler 2019, Fleurbaey and Zuber 2021, and Mongin and Pivato 2016). By a *welfare function under risk* we mean a real-valued function defined on $\mathcal{P}$ rather than X. There are two standard approaches for extending an ordinary (individual or social) welfare function W to a welfare function under risk:

- *Ex-post approach*: The welfare from a prospect $p \in \mathcal{P}$ is the expected welfare, denoted $W^{\exp}(p)$. Formally, $W^{\exp}(p) = \mathbb{E}_p(W)$. We call $W^{\exp}$ the *expectational extension* of W.

[29] This principle (defined for $X \subseteq \mathbb{R}^n$) says this: transferring a fixed material amount from a materially better-off to a materially worse-off is a social improvement. Formally, for all situations $x, y \in X$, persons i, j, and amounts $\Delta > 0$, if $y_i = x_i - \Delta$, $y_j = x_j + \Delta$, $y_i > y_j$, and $y_k = x_k$ for all other persons k, then $W(y) > W(x)$.

- *Ex-ante approach*: The welfare from a prospect $p \in \mathcal{P}$ is the equivalent welfare, denoted $W^{\text{equ}}(p)$ or more explicitly '$W^{\text{equ},\succeq}(p)$'. Formally, $W^{\text{equ}}(p) = W(x_p)$, where $x_p \in X$ is as good as p w.r.t. a given prospect order $\succeq$, i.e., $x_p \sim p$.[30] We call W^{equ} the *extension by equivalence* of W.

W^{exp} captures the welfare that will later occur in expectation, while W^{equ} captures the welfare that is initially as good as the prospect. If W is a VNM representation U of the given prospect order, then two extensions coincide: $U^{\text{exp}} = U^{\text{equ}}$. If W is instead our welfare function revealed by a (regular and broad-ranging) prospect order $\succeq$, so that $W = \log(\rho U + 1)/\rho$, where ρ is the revealed intrinsic risk attitude and U is the normalised VNM representation, then we have $W^{\text{equ}} = \log(\rho U^{\text{exp}} + 1)/\rho$. So, W^{equ} is given by the same formula as the riskless welfare function W, except that U is now replaced with its extension U^{exp} $(= U^{\text{equ}})$.

Let $\mathbf{W} : \mathcal{P} \to \mathbb{R}$ be the *social* welfare function under risk. It induces a social prospect order $\succeq_{\mathbf{W}}$. How should $\mathbf{W}$ be defined? We limit attention to utilitarianism, setting prioritarianism aside. As in the riskless case, we distinguish between W-based and VNM-based approaches, depending on how individual welfare is measured. The risky case requires a second distinction: while ex-post utilitarians maximise the total expected welfare of individuals, ex-ante utilitarians maximise the total equivalent welfare of individuals. This leads to several versions of utilitarianism under risk:

Ex-ante W-Utilitarianism: $\mathbf{W} = \sum_i W_i^{\text{equ}}$ for the welfare functions W_i revealed by the individual orders $\succeq_i$.

Ex-post W-Utilitarianism: $\mathbf{W} = \sum_i W_i^{\text{exp}} = (\sum_i W_i)^{\text{exp}}$ for the welfare functions W_i revealed by the orders $\succeq_i$.

VNM-Utilitarianism: $\mathbf{W} = \sum_i U_i^{\text{equ}} = \sum_i U_i^{\text{exp}} = (\sum_i U_i)^{\text{exp}}$ for some VNM representations U_i of the orders $\succeq_i$.

There is only one kind of VNM-Utilitarianism, since the ex-ante and ex-post approaches coincide, as $U_i^{\text{equ}} = U_i^{\text{exp}}$. This is of course an artifact of a distorted measurement of individual welfare, which neutralises the fundamental ethical distinction between ex-ante and ex-post.

Ex-post and ex-ante W-utilitarians handle risk differently. The former ignore people's risk attitudes, effectively imposing risk neutrality by taking people's expected welfare. The latter react to people's risk attitudes, by extending people's welfare functions according to people's own risk attitudes.

The axiomatic approach helps again. Here are three conditions:

[30] W^{exp} is defined if each $p \in \mathcal{P}$ has a certainty equivalent x_p and $W(x_p)$ is the same for all choices of x_p. This is guaranteed to be the case if $\succeq$ is regular and W is well-behaved (or at least compatible with $\succeq$). Regularity of $\succeq$ guarantees existence of x_p (Lemma 7), and additional well-behavedness of W guarantees that $W(x_p)$ does not depend on the choice of x_p.

Pareto Principle: For any prospects $p, q \in \mathcal{P}$, if all persons i have $p \succeq_i q$, then $\mathbf{W}(p) \geq \mathbf{W}(q)$, and if moreover one or more persons i have $x \succ_i y$, then $\mathbf{W}(p) > \mathbf{W}(q)$.

Social Rationality: The social prospect order $\succeq_{\mathbf{W}}$ is an expected-utility order, i.e., there exists a utility function $U : X \to \mathbb{R}$ such that $p \succeq_{\mathbf{W}} q \Leftrightarrow U^{\exp}(p) \geq U^{\exp}(q)$ for all prospects $p, q \in \mathcal{P}$.

Social (Intrinsic) Risk Neutrality: $\mathbf{W}$ is intrinsic risk neutral.

Social CIRA: $\mathbf{W}$ satisfies CIRA.

What does it mean for $\mathbf{W}$ to be intrinsic risk neutral or to satisfy CIRA? Initially, these two intrinsic-risk-attitudinal properties were not defined as properties of a welfare function under risk, such as $\mathbf{W}$, but of a prospect order $\succeq$ relative to a riskless welfare function W. Yet it is perfectly possible to regard intrinsic-risk-attitudinal properties as properties of a welfare function under risk, such as $\mathbf{W}$, because $\mathbf{W}$ induces a prospect order $\succeq = \succeq_{\mathbf{W}}$ and a riskless welfare function $W = \mathbf{W}|_X$.

The different versions of utilitarianism have the following properties:

Fact 4: (a) *Ex-ante W-Utilitarianism satisfies the Pareto Principle.*

(b) *Ex-post W-Utilitarianism satisfies Social Rationality and Social Risk Neutrality.*

(c) *VNM-Utilitarianism satisfies all three of these conditions, and is equivalent to this triple assuming diversity.*

The equivalence in part (c) is a version of Harsanyi's (1955) famous 'utilitarian theorem'.[31] By *diversity* we mean that, for any options $x_1, \ldots, x_n \in X$, there exists another option $x \in X$ such that $x \sim_1 x_1, x \sim_2 x_2, \ldots, x \sim_n x_n$. Diversity holds for instance in the above-mentioned 'vector model', in which social situations in X are vectors of individual situations, so that we can define x as the social situation that consists of person 1's situation in x_1, with person 2's situation in x_2, and so on.

Which version of utilitarianism is most compelling after all? While we reject VNM-Utilitarianism for its inappropriate measurement of individual welfare, the choice between

[31] His theorem achieves a conclusion closely related to VNM-Utilitarianism, but based only on a Pareto condition and a social-rationality condition, without a risk-attitudinal condition. Our result includes the third condition because our conclusion – VNM-Utilitarianism – is stronger than his conclusion, by describing a social *function* rather than *order* on $\mathcal{P}$. His social VNM order $\succeq$ is compatible with many social welfare functions $\mathbf{W}$ – all functions $\mathbf{W}$ such that $\succeq_{\mathbf{W}} = \succeq$. For sure, the most common representations $\mathbf{W}$ of his $\succeq$ are expected-utility representations. *Such* functions $\mathbf{W}$ are intrinsic risk neutral, and turn out to be writable as $\sum_i U_i^{\exp}$ for some individual VNM functions U_i. But, as Weymark (1991) insists, there are other representations of the social welfare order, such as the multiplicative representation $\prod_i e^{U_i^{\exp}}$, obtained by applying a (strictly increasing) exponential transformation to $\sum_i U_i^{\exp}$. The social function $\prod_i e^{U_i^{\exp}}$ is ordinally equivalent to $\sum_i U_i^{\exp}$, but no longer intrinsic risk neutral. Harsanyi's framework and axioms include nothing through which we could pin down a social welfare *function* and a social intrinsic risk attitude.

the ex-ante and ex-post versions of W-Utilitarianism is hard. Fact 4 shows what is at stake: Ex-ante Utilitarianism safeguards the Pareto Principle while sacrificing Social Rationality, whereas Ex-Post Utilitarianism does the reverse. But Ex-post Utilitarianism has a second notable property, which it shares with VNM-Utilitarianism: Social Risk Neutrality. Is this property a virtue or a vice? This is again debatable. On the one hand, some utilitarians might like this property, since risk neutrality is perhaps the 'most utilitarian' risk attitude, insofar as risk neutrality implies focusing on the expected welfare, and hence aggregating welfare *additively* across possible outcomes, just as utilitarians aggregate welfare *additively* across people. On the other hand, Social Risk Neutrality seems questionable if the individuals are, say, all risk averse. Instead of insisting on risk neutrality, one might give society a risk attitude that matches the average risk attitude in the group. Society is then *risk-impartial* rather than risk-neutral: it adopts people's (average) risk attitude rather than imposing risk neutrality.

This idea leads to what we will call *Risk-impartial Utilitarianism*: the theory that evaluates riskfree situations following the sum-total individual welfare (as usual) and extends this evaluation to risky prospects by applying people's (average) risk attitude. How can this be formalised? As mentioned, a riskfree welfare function $W : X \to \mathbb{R}$ can be extended expectationally to a welfare function under risk $W^{\text{exp}} : \mathcal{P} \to \mathbb{R}$. This expectational extension is implicitly risk-neutral. Let us see how to extend W by following other risk attitudes. Note first that any welfare function W and any intrinsic risk attitude $\rho \in \mathbb{R}$ jointly generate a prospect order $\succeq$, interpreted as the prospect order of someone with welfare function W and risk attitude ρ, and defined as the prospect order that is VNM represented by the utility function U given by $W = \log(\rho U + 1)/\rho$, or equivalently by $U = (e^{\rho W} - 1)/\rho$ (where as usual $U = W$ if $\rho = 0$). This prospect order is the unique prospect order with (constant) intrinsic risk attitude ρ w.r.t. W (provided that W, and thus U and $\succeq$, are regular). This follows from Proposition 2 and Corollary 1. The *extension of W via risk attitude ρ*, denoted W^{ρ}, is the extension by equivalence W^{equ} relative to the prospect order $\succeq$ generated by W and the risk attitude ρ. In the risk-neutral case $\rho = 0$, this extension is simply the expectational extension: $W^{\rho} = W^{\text{exp}}$. This is because W then equals the utility function U generated by W and $\rho = 0$. In general, W^{ρ} incorporates the risk attitude ρ. We can now define our risk-impartial version of utilitarianism:

Risk-impartial Utilitarianism: W is the extension W^{ρ} of the welfare function $W = \sum_i W_i$ via the average risk attitude $\rho = \frac{1}{n}\sum_i \rho_i$, where W_i and ρ_i are the welfare function and intrinsic risk attitude revealed by i's prospect order $\succeq_i$.

Risk-impartial Utilitarianism is a version of Ex-Post W-Utilitarianism with an impartial rather than neutral attitude to risk. Technically, the following condition is now met:

Risk Impartiality: $\mathbf{W}$ has a (social) intrinsic risk attitude $\rho_{\mathbf{W}}$ equal to the average individual intrinsic risk attitude $\frac{1}{n}\sum_{i=1}^{n}\rho_i$.[32]

Fact 5: *Risk-impartial Utilitarianism satisfies Social Rationality, Social CIRA, and Risk Impartiality.*

While we have focused on utilitarianism here, one could analogously define ex-ante, ex-post, and risk-impartial versions of prioritarianism and of many other theories of social welfare.[33]

There is an intriguing duality between social aversions to risk and to inequality. While the former is a concern about how welfare *aggregated across people* varies across possible outcomes, the latter is a concern about how welfare *aggregated across possible outcomes* varies across people. Social welfare theorists address this duality in different versions and models, exploring whether or not the goals of reducing social risk and reducing inequality stand in a conflict; see in particular Chambers (2012) and Chambers and Echenique (2012). Future research could recast the duality using our tools, including our notions of welfare and (intrinsic) risk attitude.

5.3 On the methodology for social welfare analysis

Taking a step backwards, what has been our framework to analyse social welfare, and why do we take this approach?

From ordinal observables to functional individual and social welfare

Our individual inputs are *ordinal*, given by $\succeq_1,\ldots,\succeq_n$. This makes them observable in the standard economic sense. By contrast, our social output is *functional*, given by W (riskfree case) or $\mathbf{W}$ (risky case). This functionality poses no observability problem, since outputs are not supposed to be *observed*, but to be *derived or constructed*. Still, one could reinterpret our framework such that social welfare is ordinal, given not by W (riskfree case) or $\mathbf{W}$ (risky case) but by the order $\succeq_W$ or $\succeq_{\mathbf{W}}$. Then the social function W or $\mathbf{W}$ would be no more than one of many possible numerical representations of the social order. And the various theories of social welfare, such as W-Utilitarianism and VNM-Utilitarianism, would be theories about the social welfare *order*. There is a drawback: one could no longer state conditions on social welfare that are inherently functional, i.e.,

[32] Here we treat an intrinsic risk attitude as a property of a welfare function under risk $\mathbf{W}$ rather than (as in the original definition) a property of a prospect order $\succeq$ w.r.t. a riskless welfare function W. As discussed before, it is perfectly possible to attribute intrinsic-risk-attitudinal properties to $\mathbf{W}$, because $\mathbf{W}$ induces a prospect order $\succeq=\succeq_{\mathbf{W}}$ and a riskless welfare function $W=\mathbf{W}|_X$. The intrinsic risk attitude $\rho_{\mathbf{W}}$ of $\mathbf{W}$ is simply the intrinsic risk attitude $\rho_{\succeq,W}$ of $\succeq=\succeq_{\mathbf{W}}$ w.r.t. $W=\mathbf{W}|_X$; and $\rho_{\mathbf{W}}$ exists just in case $\rho_{\succeq,W}$ is defined, i.e., $\succeq$ is regular and W is well-behaved.

[33] Risk-impartiality is partly related to Dietrich and Jabarian's (2022) notion of risk-impartial preferences of an agent who faces normative/moral uncertainty about how to evaluate risky prospects (and possibly about which risk attitude to adopt).

depend on the specific social function $\mathbf{W}$ or W rather than only on the social order $\succeq_W$ or $\succeq_{\mathbf{W}}$. One could invoke neither Social Risk Neutrality, nor Social CIRA, nor the social (intrinsic) risk attitude, all of which depend on the function $\mathbf{W}$. One could only state ordinal properties, such as the Pareto Principle and Social Rationality. This reduces the conceptual machinery for comparing ethical theories.

It is interesting to compare our framework with three established frameworks for social welfare analysis. In Amartya Sen's (1970) framework of 'social welfare functionals', individual welfare *functions* are aggregated into a social welfare *order* – the exact reverse from our setup. In Kenneth Arrow's (1951) framework, individual *orders* are aggregated into a social *order* – a purely ordinal setup. Finally, in Bergson-Samuelson's framework (Samuelson 1947; cf. Adler 2025 for a refinement), individual *functions or orders* (depending on the rendition[34]) are aggregated into a social *order*. Surprisingly, the literature has avoided a functional rendition of social welfare, despite the reasons given above. The general worry that 'functions are unobservable' might have been projected on the social level, where it does not belong, as we argued.

Ordinal or functional social welfare: a real difference?

In the case of a social welfare analysis under risk, one might at first doubt that there is a real difference between defining social welfare by a function $\mathbf{W}$ or an order $\succeq$. The reason for doubt is that a social order $\succeq$ would seem to give rise to a unique social welfare function W via Theorem 1 – just as the individual orders $\succeq_i$ reveal functions W_i – where that function W then extends to risky prospects via $\mathbf{W} = W^{\text{equ}}$. This move from ordinal to functional social welfare is however problematic. First, for a social order $\succeq$ to reveal a welfare function W via Theorem 1, $\succeq$ must be regular and broad-ranging. Regularity implies VNM-rationality. Yet social rationality cannot be taken for granted, as the counterexample of Ex-ante W-Utilitarianism shows. Second, even if $\succeq$ is regular and broad-ranging, it is not clear that our three principles – CIRA, Full-range and Normalisation – still make sense for social rather than individual welfare. Full-range might be defensible. CIRA might be too, at least if we are ready to treat society as a rational agent with a coherent attitude to risk. But Normalisation is questionable. One part of Normalisation is still plausible: it is natural to require social welfare to be 0 at the reference point, since every individual has welfare 0 there by assumption. As for the other part of Normalisation, one may doubt that marginal social welfare should be of size 1 at the reference point. The individuals have marginal welfare size 1 by assumption, and so marginal *social* welfare should be much larger than 1, since a change of situation presumably has a much larger effect on social welfare than on a single individual's welfare. For instance, if social welfare is given by the (W-)utilitarian formula $W = \sum_i W_i$, then

[34] This framework underwent some shifts in interpretation over time. Throughout, the social welfare level in a situation is modelled as a function of the individual welfare levels in that situation. While social welfare levels are commonly interpreted ordinally (as merely representing a social order), the interpretation of the individual welfare levels fluctuates between ordinal interpretations and cardinal interpretations with full interpersonal comparability. See Fleurbaey and Mongin (2005) for details.

marginal social welfare is $W' = \sum_i W'_i$, which is presumably much larger in size than each W'_i. One might respond: who cares? Normalisation sets a scale, and scales are arbitrary. Perhaps. Another response might be to apply a tailored version of Normalisation to social level – a version that sets the marginal social welfare size at the reference point to a larger value than 1. One could then obtain a social welfare function W via a version of Theorem 1 with the modified variant of Normalisation and a suitably adapted formula for W. But *which* marginal social welfare size should be required? This will inevitably depend on one's social welfare theory – utilitarians and egalitarians will disagree on the effect of situation changes on social welfare. In sum, while it may be possible to get from a VNM-rational social prospect order $\succeq$ to a social welfare function (by imposing CIRA, Full-range, and Normalisation of a variant of it), this comes at a loss of generality, as it presupposes social rationality as well as a potentially problematic scaling of social welfare.

Single-profile or multi-profile approach?

Besides the distinction between ordinal and functional welfare (for individuals and society), one commonly distinguishes between single-profile and multi-profile approaches. Our framework is single-profile: it considers a single configuration of individual and social welfare, just like Bergson-Samuelson's framework. By contrast, Arrow's and Sen's frameworks are multi-profile: they consider an entire domain of *possible* individual welfare profiles, and capture a social welfare theory by an aggregation function that maps each such profile to a corresponding social output. There is nothing wrong with multi-profile frameworks. We have limited ourselves to a single profile merely because all conditions that we happened to consider – such as the Pareto or Pigou-Dalton Principle – are intra-profile conditions. A multi-profile extension of the framework is due once we wish to study inter-profile conditions on social welfare, such as independence-type conditions.

6 Partly measurable welfare

The extent to which welfare is measurable based on ordinal evidence has been addressed by many authors in many settings. For recent formal or philosophical work on measurement scales, see Morreau and Weymark (2016), Bykvist (2021) and Nebel (2021, 2022, 2024a). Measurement scales always depend on theoretic hypotheses. In our setting, welfare is *fully* measurable if we impose CIRA, Full-range and Normalisation. To what extent is welfare measurable if we drop some of these hypotheses? This is our present question.

First, let us summarise our results on the question. Some notation will help. Given a prospect order $\succeq$, which we take to be regular and broad-ranging, let $\mathbb{U}_{\succeq}$ be the set of VNM utility representations U of $\succeq$, and U_{NOR} ($\in \mathbb{U}_{\succeq}$) the normalised one. For any $U \in \mathbb{U}_{\succeq}$ with $0 \in Rg(U)$, let $\rho_U \in \mathbb{R}$ be the intrinsic risk attitude ρ induced by U via the usual formula, given in Proposition 4 and (for normalised U) in Theorem 1. Here is our

summary:[35]

Summary of results: *Given a regular and broad-ranging prospect order $\succeq$ and a well-behaved welfare function $W : X \to \mathbb{R}$,*

(1) *W satisfies Intrinsic Risk Neutrality if and only if $W = U$ for some $U \in \mathbb{U}_{\succeq}$ (Proposition 1),*

(2) *W satisfies Intrinsic Risk Neutrality and Normalisation if and only if $W = U_{NOR}$ (a corollary of Proposition 1),*

(3) *W satisfies CIRA if and only if $W = \log(\rho U + 1)/\rho$ for some $U \in \mathbb{U}_{\succeq}$ and $\rho \in \mathbb{R}$ with $\rho U + 1 > 0$ (Proposition 2),*

(4) *W satisfies CIRA and Normalisation if and only if $W = \log(\rho U_{NOR} + 1)/\rho$ for some $\rho \in \mathbb{R}$ with $\rho U_{NOR} + 1 > 0$ (Proposition 9 in the appendix),*

(5) *W satisfies CIRA and Full-range if and only if $W = \log(\rho_U U + 1)/\rho_U$ for some $U \in \mathbb{U}_{\succeq}$ with $0 \in Rg(U)$ (Proposition 4),*

(6) *W satisfies CIRA, Full-range, and Normalisation if and only if $W = \log(\rho_{U_{NOR}} U_{NOR} + 1)/\rho_{U_{NOR}}$ (Theorem 1).*

What does this imply for the *measurability type* of welfare in these six cases? Three cases are easy to treat. In case 1, welfare is *cardinally measurable*, i.e., unique up to increasing affine transformation. In cases 2 and 6, welfare is *fully measurable*, i.e., unique. But we cannot limit attention to these three cases. For one, the cases 1 and 2 are of little interest here, since they rest on Intrinsic Risk Neutrality, which (we have argued) should be weakened to CIRA. For another, the case 6 makes the extra hypotheses of Normalisation and Full-range, which one might want to avoid.

This leaves us with the open cases 3, 4 and 5. Of what measurability type is welfare there? In case 5, the answer is surprisingly simple: welfare is *cardinally measurable*, just as in case 2. To state this result formally, let us first clarify the notion of measurability types in general. Consider a set $\mathbb{W}$ of *admissible welfare functions* $W : X \to \mathbb{R}$, interpreted as the set of welfare functions that satisfy one's hypotheses given $\succeq$. For instance, in case 5, $\mathbb{W}$ is the set of (well-behaved) welfare functions satisfying CIRA and Normalisation, or equivalently

$$\mathbb{W} = \{\log(\rho_U U + 1)/\rho_U : U \in \mathbb{U}_{\succeq} \text{ s.t. } 0 \in Rg(U)\}.$$

Measurability types, such cardinal measurability, are ultimately features of the set $\mathbb{W}$ of permissible welfare function. In particular, welfare is called

- *fully measurable* if it is unique, i.e., $\mathbb{W} = \{W\}$ for some welfare function W,
- *ordinally measurable* if it is unique up to increasing transformation, i.e., $\mathbb{W} = \{\phi \circ W : \phi \text{ is a strictly increasing transformation}\}$ for some (hence, all) $W \in \mathbb{W}$,

[35] Some of the six results in the summary hold also without $\succeq$ being regular and/or without $\succeq$ being broad-ranging.

- *cardinally measurable* if it is unique up to increasing affine transformation, i.e., $\mathbb{W} = \{aW + b : a > 0, b \in \mathbb{R}\}$ for some (hence, all) $W \in \mathbb{W}$.

The philosophical relevance of a measurability type lies in the fact that certain welfare properties become *significant*, i.e., independent of the (arbitrary) choice of welfare function from $\mathbb{W}$. In particular:

- *Ordinal measurability makes ordinal comparisons significant*: whether the welfare at x is higher, lower, or the same as the welfare at y (for $x, y \in X$) is independent of the choice of $W \in \mathbb{W}$. One can thus meaningfully say that some option gives more welfare than another, but not (for instance) that it gives much more welfare, or twice as much welfare.
- *Cardinal measurability makes ratios of differences significant*: the ratio $\frac{W(x)-W(y)}{W(x')-W(y')}$ (for $x, y, x', y' \in X$) is independent of the choice of welfare function $W \in \mathbb{W}$.[36] One can thus meaningfully say that a certain welfare gain is twice as large as another one, but not (for instance) that this gain exceeds 3, or that it is large.
- *Full measurability makes all welfare features significant.* For instance, one can meaningfully say that welfare is 7 at x, or that welfare decreases by 2 from x to y (for $x, y \in X$).

Informally, (in)significant features of a welfare function $W \in \mathbb{W}$ are features that *can (not) be taken seriously*.[37] Since the measurability type of welfare is determined by the set $\mathbb{W}$ of admissible welfare functions, it is indirectly determined by the prospect order $\succeq$ combined with one's welfare hypotheses (such as CIRA). We can thus say that welfare is of some measurability type '*given* $\succeq$ *and certain welfare hypotheses*' – which stands for '*given* $\mathbb{W}$', with $\mathbb{W}$ defined as the set of welfare functions that satisfy these hypotheses relative to $\succeq$'.

Here is the surprising fact about case 5, in which CIRA and Full-range are the only welfare hypotheses:

Proposition 5 *Given a regular and broad-ranging prospect order $\succeq$ and the welfare hypotheses of CIRA and Full-range (and well-behavedness), welfare is cardinally measurable.*

The cardinal measurability of welfare is an unexpected point of convergence between VNM utility and our welfare measure (without Normalisation). Although these two welfare measures are very different, welfare is both times cardinally measurable. So,

[36] More precisely, what is independent of W is the ratio's value if it is well-defined (i.e., if $W(x') - W(y') \neq 0$) and the ratio's well-definedness (i.e., the fact of whether $W(x') - W(y') \neq 0$).

[37] More fundamentally, one may distinguish between two possible interpretations of (in)significance. Under a *metaphysical* interpretation, insignificant features of W do not actually correspond to any fact about welfare in the world. Under an *epistemological* interpretation, insignificant features of W do not correspond to a *knowable* fact about welfare in the world: the available evidence does not allow one to tell whether the agent's true welfare has the corresponding feature. In short, insignificance is either lack of meaning or empirical underdetermination. Arguably, the epistemological interpretation if often more appropriate. Under this interpretation, limited measurability of welfare is a matter of limited knowledge about welfare, not of principled meaninglessness of certain propositions about welfare.

under both approaches, we can meaningfully talk about ratios of welfare differences, and hence about comparisons of welfare differences, while welfare levels and welfare changes are insignificant (i.e., empirically underdetermined or even metaphysically meaningless, as explained in footnote 37). What should we make of the fact that the two welfare measures have difference ratios and difference comparisons that are *significant, but distinct*? What matters is not only *that* difference ratios and difference comparisons are significant, but also *what* they signify. They signify something different! For our measure W, they arguably signify facts about the agent's real welfare (or preference, under the preference interpretation). An inequality $W(y) - W(x) > W(y') - W(x')$ means that the change from x to y is better than that from x' to y'. By contrast, for a VNM utility function U, difference ratios and difference comparisons do not signify facts about welfare alone. Instead they signify some complicated hybrid facts about the interplay of welfare and risk attitude. An inequality $\frac{U(y)-U(x)}{U(y')-U(x')} > 1$ does not mean that the change from x to y is better than that from x' to y'. This point is of course controversial, and far from new. Already von Neumann and Morgenstern (1944) warned against hasty interpretations of ratios or comparisons of utility differences in terms of preference strength. On the debate, see Ellsberg (1954), Baumol (1958), and Fishburn (1989, sec. 2.1).

Now suppose that we no longer wish to impose Full-range. Our only welfare hypothesis is CIRA (besides well-behavedness). This is case 3 above. Here, welfare is no longer cardinally measurable. Still, 'half' of cardinal measurability still holds. Let us call welfare *semi-cardinally measurable* if the set $\mathbb{W}$ of admissible welfare functions is closed under increasing affine transformation, i.e., if $W \in \mathbb{W} \Rightarrow aW + b \in \mathbb{W}$ for all $a > 0$ and $b \in \mathbb{R}$. Semi-cardinal measurability is less demanding than cardinal measurability, because the set equality '$\mathbb{W} = \{aW + b : W \in \mathbb{W}\}$' defining cardinal measurability is weakened to the set inclusion '$\mathbb{W} \supseteq \{aW + b : W \in \mathbb{W}\}$'.[38]

Proposition 6 *Given a regular prospect order $\succeq$ and the welfare hypothesis of CIRA (and well-behavedness), welfare is semi-cardinally measurable.*

Finally, what happens in case 4, where our welfare hypotheses are CIRA and Normalisation? Here welfare is not even semi-cardinally measurable, since normalised welfare functions become non-normalised when transformed in an affine way. Giving complete characterisations of the measurability types in cases 3 and 4 is an exercise beyond the scope of this paper. (Note that Proposition 6 gives only a partial characterisation for case 4.) These two measurability types are non-standard, since there is no set of transformations up to which welfare is unique.[39]

[38] Also, the quantification 'for some (hence, all)' is weakened into 'for all'. So, $\mathbb{W}$ can be empty.

[39] That is, $\mathbb{W}$ is not writable as $\{\phi \circ W : \phi \in \Phi\}$ for some $W \in \mathbb{W}$ and some set Φ of transformations $\phi : \mathbb{R} \to \mathbb{R}$ (where $\phi, \psi \in \Phi \Rightarrow \phi \circ \psi \in \Phi$ and where each $\phi \in \Phi$ is bijective with $\phi^{-1} \in \Phi$). By contrast, full, ordinal, and cardinal measurability are *standard* measurability types, with Φ containing only the identity transformation, all increasing transformations, resp. all increasing affine transformations. Non-standard measurability types, such as those in cases 3 and 4, could for instance be characterised by identifying the *significant* welfare features derivable from $\mathbb{W}$.

It would be interesting to study *social* welfare based on non-fully measurable individual welfare. Consider again our social-welfare framework (Section 5), but without imposing Normalisation on individual welfare. Then individual welfare is only cardinally measurable, so that the functions W_i are only unique up to increasing affine transformation (Proposition 5). In result, the W-utilitarian social welfare function $\sum_i W_i$ is non-unique, as is the induced social welfare order. But other social welfare functions still lead to a unique social order, notably the Nash social welfare function.[40]

7 Discussion and generalisation

In this section, we discuss the plausibility of our three welfare hypotheses, and present generalisations of them and of our theorem. We start with Full-range (Section 7.1), followed by Normalisation (Section 7.2), CIRA (Section 7.3), and the generalised theorem (Section 7.4). We finally make a case for the original versions of our hypotheses (Section 7.5).

7.1 Full-range discussed and generalised

This and the following subsections discuss possible objections to the three hypotheses, which will lead us to generalise these hypotheses and the theorem. We begin with Full-range, followed in later subsections by the other hypotheses and the theorem.

As measurement theorists will notice, Full-Range is not only a richness condition on X (that forces X to contain arbitrarily good or bad situations) but also implicitly a condition on the choice of measurement scale for welfare: that scale should have the range $\mathbb{R}$, so that all real numbers are *meaningful* as welfare levels. Scales with smaller range are also imaginable. For instance, a scale with range $\mathbb{R}$ can be transformed exponentially into one with range $\mathbb{R}_+$, by replacing ('relabelling') any welfare level $w \in \mathbb{R}$ with $e^w \in \mathbb{R}_+$. To allow many measurement scales, we fix a non-empty open interval $D \subseteq \mathbb{R}$ of *meaningful* welfare levels, e.g., $D = \mathbb{R}$ or $D = (0, \infty)$ or $D = (0, 1)$, and impose the following hypothesis (which reduces to Full-range if $D = \mathbb{R}$):[41]

Full-range$_D$: There are situations of arbitrary quality in D, i.e., $\{W(x) : x \in X\} = D$.

7.2 Normalisation discussed and generalised

This condition sets welfare to 0 and marginal welfare size to 1, at a given reference point $\overline{x}$, for instance a poverty level. More generally, one can fix numbers $r \in D$ and $s > 0$, and place the following requirement, where we call a function from X to $\mathbb{R}$ (r, s)-*normalised* if at the reference point $\overline{x}$ it takes the value r and has a derivative of size s:

[40] The latter is given by $\prod_i (W_i - W_i(\bar{x}))^{1/n}$, and is defined only for situations $x \in X$ in which $W_i(x) \geq W_i(\bar{x})$ for all i.

[41] A metaphysical background assumption is that uncountably infinitely many welfare states exist.

Normalisation$_{r,s}$: The welfare function W is (r, s)-normalised.

What speaks for requiring Normalisation$_{r,s}$ instead of Normalisation, the special case with $r = 0$ and $s = 1$? Normalisation is questionable when one makes interpersonal comparisons of welfare, since Normalisation treats everyone as having the same welfare (and marginal welfare size) at $\overline{x}$. Nothing is wrong with assuming Normalisation for for a *given* person – this just requires choosing a measurement scale on which '0' and '1' have particular meanings adapted to that person (and such a scale always exists, as noted earlier). But assuming Normalisation for many persons simultaneously leads to questionable welfare comparisons at $\overline{x}$, since '0' and '1' can be given just one meaning at once. By contrast, Normalisation$_{r,s}$ works even in an interpersonal context, since r and s can be set differently for different persons.

In some contexts, Normalisation *is* defensible. Why? First of all, recall that choices of normalisation are an old problem in welfare economics, although it is usually raised for VNM utility rather than welfare. Already Harsanyi was bothered by the sensitivity of interpersonal utility comparisons and total utility in society to normalisation choices. Different proposals exist. Some fix utility at two reference outcomes, e.g., to 0 and 1 (e.g., Isbell 1959, Segal 2000, Adler 2012, 2016). Others fix the minimal and maximal utility (Karni and Weymark 2024). Fleurbaey and Zuber (2021) instead fix utility and marginal utility at a reference outcome. Our Normalisation follows their approach, applying it to welfare rather than VNM utility.

In fact, Normalisation is appropriate in two contexts:

1. Locally objective welfare: One often distinguishes 'objective' from 'subjective' notions of welfare (Fleurbaey and Blanchet 2013). Objective welfare is determined by 'external' or 'objective' features like wealth or consumption levels, subjective welfare by 'internal' or 'subjective' features like tastes for (or happiness from) wealth or consumption. Assuming the situations in X are *objective* situations, objective welfare depends on the situation alone, while subjective welfare also depends on subjective features such as tastes. A notion of welfare can be of hybrid objective-subjective nature, so that welfare depends partly on objective and partly on subjective factors, where the extent of objectivity can vary across situations. One can interpret Normalisation as requiring welfare to be *objective (at least) at the reference point* $\overline{x}$: at $\overline{x}$, the person's welfare and marginal welfare size are objectively fixed. For instance, a situation of misery $\overline{x}$ might lead *objectively* to a certain (low) well-being and (high) marginal welfare size, whereas non-miserable situations give room for subjective tastes to influence welfare. The slogan is: in situations of objective misery everyone is alike welfare-wise. Subjective features, such as whether one prefers Beethoven's or Bach's music, influence welfare only outside situations of extreme misery, when basic needs are satisfied. On this assumption, Normalisation is justified, subject to making the scaling convention that 0 (1) stands for the universal welfare (marginal welfare size) at $\overline{x}$. Similar ideas of locally objective welfare are popular in the theory of fair allocation, as mentioned in Section 4.

In fact, it suffices that welfare be *effectively* locally objective at $\overline{x}$: at $\overline{x}$, welfare can

still depend partly on subjective features as long as these features coincide for everyone at $\overline{x}$ – so that welfare at $\overline{x}$ depends on *subjective but universal* features. So, welfare can depend on subjective tastes, as long as everyone dislikes the situation of misery $\overline{x}$ *equally*, i.e., 'needs' subjectively the basic needs equally. This makes welfare effectively locally objective. More precisely, welfare is then locally *intersubjective*, as we say.[42]

The general idea is that the objective circumstances take over at $\overline{x}$, making subjective differences inexistent (true local objectivity) or irrelevant to well-being (local intersubjectivity). The plausibility of this idea arguably depends partly on the notion of situation in X. Mere wealth levels are perhaps too uninformative 'situations' for the welfare level to become objective at some 'poverty point' $\overline{x}$. Things might change if situations are detailed consumption vectors, or entire 'lives', or Sen-type functioning vectors. The more information is packed into situations – perhaps including quasi-subjective features – the less room is left for subjectivity in welfare assessments.

Stigler and Becker's (1977) thesis 'De gustibus non est disputandis' and Sen's (1985) programme of evaluating fine-grained functioning vectors are two very different attempts at objective evaluations through refining the description of situations.

2. Egalitarian normalisation: Suppose the question is not what welfare the persons *have* intrinsically, but what welfare they should be *treated* as having in order to lead to egalitarian social redistributions. More precisely, assume we wish to scale individual welfare such that a social planner who maximises total individual welfare will redistribute goods from the poor to the rich. Under reasonable assumptions, Normalisation achieves precisely this! Why? Assume $X \subseteq \mathbb{R}$ and the reference point $\overline{x} \in X$ represents a 'poverty point' below/above which someone counts as poor/rich. Let social situations be vectors $(x_1, ..., x_n) \in X^n$, where x_i represents i's situation. Adopting (W-)Utilitarianism, let the social welfare in a situation $x \in X^n$ be $\sum_i W_i(x_i)$. If all W_i satisfy Normalisation and are increasing and strictly concave, then transferring resources from rich to poor persons raises social welfare $\sum_i W_i(x_i)$.[43] So, Normalisation gives utilitarianism an unexpected egalitarian appeal. This egalitarian argument for Normalisation is introduced and developed axiomatically in Fleurbaey and Zuber (2021), in a version for VNM utility instead of welfare. In their words, Normalisation leads to 'fair utilitarianism'.

7.3 CIRA discussed and generalised

CIRA requires a constant attitude to intrinsic risk, i.e., to risk in welfare rather than outcomes – a welfare-level condition that can explain well-documented violations of the outcome-level condition CARA (see Fact 2 and Proposition 8).

One can defend CIRA by deriving it from two basic assumptions on how prospects are ordered. How? In principle, the agent's prospect order could depend on the status-

[42] Even if welfare is objective (really or effectively, locally or globally), people may differ in intrinsic risk attitudes, affecting their VNM utility functions.

[43] That is, for all social situations $x, y \in X^n$ and individuals j, k, if $x_j < y_j < \overline{x} < y_k < x_k$, $y_j - x_j = x_k - y_k$, and $x_l = y_l$ for everyone else l, then $\sum_i W_i(x_i) < \sum_i W_i(y_i)$.

quo situation, e.g., an initial wealth level. To make this idea explicit, consider an entire *order structure*, i.e., a family $(\succeq_z)_{z\in X}$ containing the prospect order $\succeq_z$ held in any possible status quo $z \in X$. In a status quo $z \in X$, each lottery $p \in \mathcal{P}$ induces a *welfare-change lottery*, denoted $p_{\Delta W|z}$, defined as the finite-support lottery over $\mathbb{R}$ such that the probability of any welfare change $w \in \mathbb{R}$ is $p_{\Delta W|z}(w) = p(W = W(z)+w)$, the probability that final welfare is initial welfare plus w. An order structure $(\succeq_z)_{z\in X}$ is

(a) *difference-based* if, for all prospects $p, q, p', q' \in \mathcal{P}$ and stati quo $z, z' \in X$, if $p_{\Delta W|z} = p'_{\Delta W|z'}$ and $q_{\Delta W|z} = q'_{\Delta W|z'}$ then $p \succeq_z q \Leftrightarrow p' \succeq_{z'} q'$,

(b) *dynamically consistent* or *status-quo independent* if $\succeq_z$ is the same for each status quo $z \in X$.

Proposition 7 *If an order structure $(\succeq_z)_{z\in X}$ satisfies (a) and (b), then the prospect order $\succeq_z \equiv \succeq$ satisfies CIRA, assuming it is regular and W is well-behaved and satisfies Full-range.*

Condition (b) follows orthodox rational choice theory. Condition (a) follows Kahneman-Tversky's popular approach of modelling decisions as choices between *changes* rather than final results, but in an (arguably more plausible) version based on welfare changes rather than outcome changes. Kahneman-Tversky's idea is that real agents tend to conceptualise options in terms of changes rather than final consequences, since changes represent what 'happens' in the agent's perception.

More importantly perhaps, condition (a) requires a situation-independent and thus robust attitude to risk of losing or gaining welfare. The situation has no effect on whether the agent prefers (say) facing no welfare change for sure to facing a 50:50 lottery between gaining and losing one unit of welfare. Condition (a) is plausible to the extent that the attitude to welfare-change risk is a deeply entrenched personality trait. Some people are cautious or fearful, others playful or reckless – *by character*, independently of their situation or welfare.

If one nonetheless rejects CIRA, one can replace it with a more flexible hypothesis, with many special cases, such as cases of *decreasing* intrinsic risk aversion, or of constant *relative* intrinsic risk aversion. We define the generalised condition as requiring a constant attitude to risk in *some* welfare-based quantity, i.e., in some transformation of welfare. More precisely, given a welfare transformation, which can be any smooth function τ from the mentioned interval D of meaningful welfare levels onto $\mathbb{R}$ with $\tau' > 0$, we require this:

CIRA$_\tau$: If a prospect is modified by increasing each possible resulting transformed welfare by the same amount, then the transformed equivalent welfare increases by the same amount. Formally, $Rg(W) \subseteq D$ and, for all $\Delta > 0$ and all prospects $p \in \mathcal{P}$ with an equivalent welfare $w_p \in \mathbb{R}$, if another prospect $q \in \mathcal{P}$ with an equivalent welfare $w_q \in \mathbb{R}$ satisfies $q(\tau(W) = t + \Delta) = p(\tau(W) = t)$ for all $t \in \mathbb{R}$, then $\tau(w_q) = \tau(w_p) + \Delta$.

CIRA$_\tau$ reduces to CIRA if $\tau(w) = w$ for all $w \in D = \mathbb{R}$. If instead $\tau(w) = \log w$ for

all $w \in D = (0, \infty)$, then CIRA_τ requires constant aversion to risk in welfare *ratios*, i.e., constant *relative* intrinsic risk aversion.

7.4 The theorem generalised

Even in their generalised form, our hypotheses lead to a unique welfare measure:

Theorem 2 *Given a regular and broad-ranging prospect order $\succeq$, there exists a unique well-behaved welfare function W satisfying $CIRA_\tau$, Full-range$_D$, and Normalisation$_{r,s}$ (for given parameters τ, D, r, s), namely the function*

$$W = \tau^{-1}(\log(\rho s \tau'(r)U + 1)/\rho + \tau(r))$$

based on the normalised (unbounded) VNM representation U of $\succeq$ and the parameter

$$\rho = \begin{cases} \frac{-1}{s\tau'(r)\sup U} < 0 & \text{if } \sup U \neq \infty \\ \frac{-1}{s\tau'(r)\inf U} > 0 & \text{if } \inf \neq -\infty \\ 0 & \text{if } \sup U = \infty \text{ and } \inf U = -\infty. \end{cases}$$

Theorem 1 is a special case of Theorem 2, obtained if $D = \mathbb{R}$, $r = 0$, $s = 1$, and τ is the identity transformation, because the three hypotheses then reduce to the original ones, and the formulas reduce to those in Theorem 1.[44]

For instance, if $D = (0, \infty)$ and $\tau = \log$, so that CIRA_τ requires a constant *relative* intrinsic risk attitude, then the formula for W reduces to $W = r(\frac{\rho s}{r}U + 1)^{1/\rho}$, so that welfare a geometric function of VNM utility.

7.5 In defence of the initial form of the hypotheses

Which versions of our hypotheses is most compelling after all? Should one impose the initial versions CIRA, Full-range and Normalisation, and measure welfare via Theorem 1, or impose modified versions, and measure welfare via Theorem 2? The key difference lies in the *scale* on which welfare is measured. The initial hypotheses lead to a scale with range $\mathbb{R}$. Were one to transform this scale exponentially, so that each old welfare level $w \in \mathbb{R}$ corresponds to the new welfare level e^w, then the initial welfare function W is replaced with the new one $\hat{W} = e^W$. W and $\hat{W}$ capture the same reality, but on different scales or 'languages'. While W satisfies CIRA, Full-range and Normalisation, $\hat{W}$ satisfies CIRA_τ, Full-range$_D$ and Normalisation$_{r,s}$ with parameters $D = (0, \infty)$, $\tau = \log$, and $r = s = 1$.

The choice of scale matters greatly. For instance, a utilitarian, who maximises $\sum_i W_i$, could not simply change the scale of individual welfare exponentially, since this would imply replacing each W_i with $\hat{W}_i = e^{W_i}$, and replace the rule to maximise $\sum_i W_i$ with the (equivalent) rule to maximise $\prod_i \hat{W}_i$ – no longer a utilitarian rule.

[44] If $\rho = 0$ then the expression $\log(\rho s \tau'(r)U + 1)/\rho$ in the formula for W stands for $s\tau'(r)U$ $(= \lim_{\rho \to 0} \log(\rho s \tau'(r)U + 1)/\rho))$.

Normally, a good scale should at least ensure that the welfare measure W *permits intrapersonal comparisons of differences*: the difference $W(y) - W(x)$ is the higher, the better the change from x to y is, across all situations $x, y \in X$.

Conjecture 1: Given an observed prospect order $\succeq$, in order to derive a welfare function W that permits comparing differences intrapersonally, it is better to assume the basic hypotheses CIRA and Full-range than other variants CIRA_τ and Full-range_D, and it is as good to assume the basic hypothesis Normalisation than any variant $\text{Normalisation}_{r,s}$.

Before justifying this conjecture, note that social welfare analysis usually needs more: it needs *interpersonal* comparisons. Welfare functions W_i of persons $i = 1, \ldots, n$ *permit interpersonal comparisons of differences* if $W_i(y) - W_i(x)$ is the higher, the better the change from x to y is for person i, across all persons i and situations $x, y \in X$. Utilitarians, who maximise $\sum_i W_i$, rely on such comparisons. Prioritarians and many other egalitarians also need to compare welfare *levels*. The welfare functions W_i *permit interpersonal comparisons of levels* if $W_i(y)$ is the higher, the better off person i is in x, across all persons i and situations $x \in X$.

Conjecture 2: Given observed prospect orders $\succeq_i$ of persons $i = 1, \ldots, n$, in order to derive welfare functions W_i that permit comparing levels and differences interpersonally, one should assume CIRA and Full-range rather than other variants, and assume $\text{Normalisation}_{r_i,s_i}$ with well-calibrated person-specific parameters r_i, s_i. If using person-specific parameters is infeasible or inappropriate, then assuming Normalisation is as good as assuming any variant $\text{Normalisation}_{r,s}$.

These are two *informal* conjectures, without a formal statement or proof, since scales are informal objects here. A scale can be conceptualised as a correspondence between informal properties in the world (welfare states of a person) and numbers (formal welfare levels). Since one side of the correspondence is outside our model, the correspondence is not a formal object.

Why these conjectures? Let us sketch the argument. We begin with defending the superiority of CIRA and Full-range over their variants. The argument for this will rely only on *intrapersonal* comparisons of differences only. First, consider CIRA. CIRA can be paraphrased as follows: an indifference between a sure welfare level w and a risky welfare change from w (e.g., a 50:50 lottery of gaining or losing one unit of welfare from w) does not depend on w – i.e., the indifference holds for all w or for no w. So-restated, CIRA seems plausible *provided* that a given welfare change from w (such as: gaining one unit of welfare) can be treated as *the same gain* regardless of the level w from which it started. So, the plausibility of CIRA rests on the comparability of differences. As does our more systematic argument for CIRA in Section 7.3, where we defended CIRA based largely on the idea that the person has a robust attitude to risk of losing or gaining welfare, interpreted as a stable character trait. An analogous argument could *not* be made for CIRA_τ (assuming τ is non-affine, to ensure that CIRA_τ is not equivalent to CIRA).

Worse, an analogous argument could be made *against* CIRA_τ: CIRA_τ effectively rules out a stable attitude to risk of gaining or losing welfare, instead requiring a particular pattern of attitudinal change or 'character instability' – an implausible requirement. Just as the argument for CIRA, the argument against CIRA_τ relies on being able to meaningfully compare welfare differences.

Full-range also seems more plausible than Full-range$_D$ with $D \neq \mathbb{R}$, given that welfare differences are to be comparable, and assuming X is sufficiently rich in situations. Why? Fix a situation x_0 in X. Pick new situations $x_1, x_2, x_3, \ldots$ such that x_1 is better than x_0, and the change from x_0 to x_1 is as good as that from x_1 to x_2, and as that from x_2 to x_3, and so on. Then, since W has meaningful differences, $0 < W(x_1) - W(x_0) = W(x_2) - W(x_1) = W(x_3) - W(x_2) = \ldots$ So, $W(x_k) \to \infty$ as $k \to \infty$, and thus $\sup W = \infty$. Analogously, $\inf W = -\infty$. So, $Rg(W) = \mathbb{R}$, i.e., Full-range hods, while Full-range$_D$ with $D \neq \mathbb{R}$ is violated.

The normalisation axiom has a different status. It is crucial for *interpersonal* comparisons. If for whatever reason we know people's well-being at and around the reference point, then we can introduce person-specific level parameters r_i and sensitivity parameters s_i, and impose Normalisation$_{r_i,s_i}$ on W_i. The calibration of the sensitivity parameters s_i enables interpersonal comparability of differences – initially near the reference point, and then anywhere else via intrapersonal comparisons of differences.[45] This also implies interpersonal comparability of levels, thanks to the well-calibrated level parameters r_i.[46] Often, however, the use of person-specific parameters r_i and s_i is impossible. Either it is infeasible, since nothing is known about people except their prospect orders. Or it is inappropriate, since people *are* or *should be treated as* similar at the reference point (see Section 7.2). We must then impose a person-unspecific normalisation axiom. Whether we choose the basic axiom Normalisation or a variant Normalisation$_{r,s}$ makes no difference to interpersonal comparability.[47]

[45] By the choice of the sensitivity parameters s_i, if x is the reference point $\overline{x}$, the difference $W_i(y) - W_i(x) = \delta$ is the higher, the better the change from x to y is for person i (at least approximately, for y close to x). This interpersonal comparability of differences holds also when x is not the reference point, since *all* changes of situation that result in the same welfare difference δ for a given person i are equally good for i, by *intrapersonal* comparability of differences.

[46] How might one calibrate the r_i's? Let us begin with arbitrary parameters r_i. So far, only the parameters s_i are calibrated. Pick a situation y in which everyone is equally well off (assume it exists). First assume $W_i(y)$ is the same for all i. Then welfare levels are interpersonally comparable, because, for all situations x and persons i, $W_i(x)$ is the higher, the higher the difference $W_i(x) - W_i(y)$ is (since $W_i(y)$ is fixed), i.e., the better the change from y to x is for i (by interpersonal comparability of differences), i.e., the better off i is in x (since y is fixed). So, the r_i's are already well-calibrated. Now assume $W_i(y)$ is not the same for all i – a violation of interpersonal comparability of levels. Then recalibrate the r_i's by replacing each r_i with $r_i^* = r_i - W_i(y)$. The new conditions Normalisation$_{r_i^*,s_i}$ yield new welfare measures $W_i^* = W_i - W_i(y)$. This time, levels are interpersonally comparable, by the previous argument applied to the functions W_i^*, which take the same value (of 0) at y. So the r_i^*'s are well-calibrated.

[47] Why? CIRA and Full-range already determine the welfare functions W_i up to increasing affine transformation (Proposition 5). A normalisation condition makes the W_i's unique. The difference between two versions Normalisation$_{r,s}$ and Normalisation$_{r',s'}$ is that all W_i's change via the same increasing affine transformation. Such a transformation preserves interpersonal comparisons of levels and of differences.

To sum up, the hypotheses CIRA and Full-range seem more appropriate than their variants, and the hypothesis Normalisation is appropriate unless person-specific normalisation is possible.

8 Conclusion

We have shown that plausible working hypotheses, most importantly Constant Intrinsic Risk Attitude, allow us to operationalise the notion of welfare or intrinsic utility. Our operationalisation contrasts with the standard one in terms of VNM utility, which rests on a less plausible hypothesis than CIRA, namely Intrinsic Risk Neutrality. Our approach allows us to decompose standard utility and standard risk attitude into their two determinants: welfare and intrinsic risk attitude, both of which are indirectly observable. We have presented formal and informal arguments for adopting our hypotheses as working assumptions. But we have also analysed how welfare becomes *partially* observable if only some of our three hypotheses are adopted. In particular, if we drop the Normalisation hypothesis, welfare is cardinally measurable, just as VNM utility is cardinally measurable.

Social welfare analysis can now use a more satisfactory observable measure of personal welfare than VNM utility. It can disentangle welfare aspects from risk-attitudinal aspects, instead of mixing both aspects unrecognisably within VNM utility. So far, critics of VNM utility in social welfare analysis, such as Amartya Sen and followers, failed to operationalise their position and to address the unobservability objection.

In our treatment of social welfare, we have sketched first avenues and put forward methodological claims. A systematic analysis will need to address the *social* attitude to *intrinsic* risk. This could be achieved by aggregating people's (revealed) intrinsic risk attitudes. A new version of utilitarianism – *risk-impartial utilitarianism* – pursues this approach.

A Appendix

A.1 The generalised setup with arbitrary alternatives

The main text took the set of alternatives X to be a (non-empty open connected) subset of $\mathbb{R}^k$ for some $k \geq 1$. Our definitions, hypotheses and results continue to hold for an arbitrary non-empty set X, except for those about the classical Arrow-Pratt risk attitude (Definition 1, Fact 2, and Propositions 3 and 8). This generalisation requires generalised notions of 'regular' and 'normalised' functions on X, since derivatives are undefined in general. Here are the details, for interested readers:

Regularity generalised. In the main text, a function $W : X \to \mathbb{R}$ counted as 'regular' if it is smooth with nowhere zero derivative. In general, the set of *regular* functions is any given set $\mathcal{F}$ of functions $f : X \to \mathbb{R}$ such that for all $f \in \mathcal{F}$, the range $Rg(f) = \{f(x) :$

$x \in X\}$ is an open interval and, for each strictly increasing $\phi : Rg(f) \to \mathbb{R}$, $\phi \circ f \in \mathcal{F} \Leftrightarrow$ ϕ is smooth with $\phi' > 0$. The main text's regularity notion is a special case, as we will soon see.

Normalisation generalised. In the main text, a function $W : X \to \mathbb{R}$ counted as 'normalised' if it has value 0 and a derivative of size 1 at the reference point $\overline{x}$. In general, the set of *normalised* functions is any given set $\mathcal{N}$ of functions $f : X \to \mathbb{R}$ such that (i) for all $f \in \mathcal{N}$, $f(\overline{x}) = 0$, (ii) for all $f \in \mathcal{N}$ and all smooth transformations $\phi : Rg(f) \to \mathbb{R}$ with $\phi(0) = 0$, we have $\phi \circ f \in \mathcal{N} \Leftrightarrow \phi'(0) = 1$, (iii) each function $f \in \mathcal{F}$ is normalisable, i.e., has an increasing affine transformation in $\mathcal{N}$. Normalised functions have the right value at $\overline{x}$ by (i), and intuitively the same 'abstract derivative' at $\overline{x}$ by (ii).[48] We must also generalise the related notion of an (r, s)-*normalised* function $f : X \to \mathbb{R}$, for $r \in \mathbb{R}$ and $s > 0$. In the main text, '(r, s)-normalised' means that $f(\overline{x}) = r$ and f has a derivative of size s at $\overline{x}$. In general, it means that $f = sg + r$ for some normalised function $g \in \mathcal{N}$.

The concrete setup of the main text is indeed a special case:

Lemma 1 *The conditions on sets $\mathcal{F}$ and $\mathcal{N}$ of regular resp. normalised functions hold if (as in the main text) X is a non-empty open connected subset of $\mathbb{R}^k$ with $k \geq 1$ and*

$$\mathcal{F} = \{f : X \to \mathbb{R} : f \text{ is smooth, } f'(x) \neq \mathbf{0} \text{ for all } x \in X\}$$
$$\mathcal{N} = \{f : X \to \mathbb{R} : f(\overline{x}) = 0,\ f'(\overline{x}) \text{ exists and is of size 1}\}.$$

Proof. Let X, $\mathcal{F}$ and $\mathcal{N}$ be as in the main text. We first establish the conditions on $\mathcal{F}$ (part 1), then those on $\mathcal{N}$ (part 2).

1. Fix an $f \in \mathcal{F}$. $Rg(f)$ is an interval, since continuous images of connected sets are connected. This interval is open, because X is open and $f'(x) \neq \mathbf{0}$ for all $x \in X$. Now fix a strictly increasing $\phi : Rg(f) \to \mathbb{R}$. By basic calculus, if ϕ is smooth with $\phi' > 0$, then $\phi \circ f \in \mathcal{F}$.

Conversely, assume $\phi \circ f \in \mathcal{F}$. We must show that ϕ is smooth with $\phi' > 0$. Put $g = \phi \circ f$.

Claim 1: If $X \subseteq \mathbb{R}$ (i.e., $k = 1$), then f^{-1} exists and is smooth.

Let $X \subseteq \mathbb{R}$. As $f \in \mathcal{F}$, the derivative f' exists and is continuous and nowhere zero. So f' is everywhere positive or everywhere negative. Thus f is strictly monotonic, hence invertible. To show that $h = f^{-1}$ is smooth, we show by induction that for all $n \geq 1$ the n^{th} derivative $h^{(n)}$ exists and is a ratio $\frac{a}{b}$ of smooth functions $a, b : X \to \mathbb{R}$ with $b > 0$. First consider $n = 1$. As $f' > 0$, the function, $h' = (f^{-1})'$ exists and equals $\frac{1}{f'(h)}$, a ratio of the claimed form. Now let $n > 1$ and assume that $h^{(n-1)}$ exists and takes the claimed

[48] In (ii), $\phi \circ f$ intuitively has the same abstract derivative at $\overline{x}$ as f if and only if $\phi'(0) = 1$. Intuitive reason: $(\phi \circ f)'(\overline{x}) = \phi'(f(\overline{x}))f'(\overline{x}) = \phi'(0)f'(\overline{x})$, assuming abstract derivatives behave like ordinary ones, and (i) holds.

form, say $h^{(n-1)} = \frac{a}{b}$. By implication, $h^{(n)}$ exists and equals $\frac{a'b - b'a}{b^2}$, another ratio of the claimed form. Q.e.d..

Claim 2: If $X \subseteq \mathbb{R}$ (i.e., $k = 1$), then ϕ is smooth with $\phi' > 0$.

Assume $X \subseteq \mathbb{R}$. As $g = \phi \circ f$ and f is (by Claim 1) invertible, we have $\phi = g \circ h$, where $h = f^{-1}$. The smoothness of ϕ can be deduced from the fact that $\phi = g \circ h$ and that g and (by Claim 1) h are smooth. How? In short, ϕ' exists and equals $h'g'(h)$; so ϕ'' exists and equals $h''g'(h) + h'(g'(h))' = h''g'(h) + h'^2 g''(h)$; and so on for higher derivatives of ϕ (we skip the full inductive argument).

To see why $\phi' > 0$, fix a $w \in Rg(f)$. Pick an $x \in X$ such that $f(x) = w$. We have $g'(x) = \phi'(w)f'(x)$ since $g'(x) = (\phi \circ f)'(x) = \phi'(f(x))f'(x) = \phi'(w)f'(x)$. So, as $g'(x)$ and $f'(x)$ are non-zero and (by ordinal equivalence of f and g) of same sign, we have $\phi'(w) > 0$. Q.e.d.

Claim 3: In general, ϕ is smooth with $\phi' > 0$ (completing the proof).

Now we allow X to be multi-dimensional: X is any non-empty open connected subset of $\mathbb{R}^k$ with $k \geq 1$. Let $t \in Rg(f)$. We must show that, at t, ϕ is smooth with $\phi' > 0$. Pick an $x \in f^{-1}(t)$. Since $f'(x) \neq \mathbf{0}$, we may pick a coordinate $j \in \{1, ..., k\}$ such that $\frac{df}{dx_j}(x) \neq 0$. As f is smooth and f' is nowhere zero, there is an open interval $\tilde{X}$ containing x_j such that, for all $y \in \tilde{X}$, $(x_1, ..., x_{j-1}, y, x_{j+1}, ..., x_k) \in X$ and $\frac{df}{dx_j}(x_1, ..., x_{j-1}, y, x_{j+1}, ..., x_k) \neq 0$. Consider f as a function of the j^{th} coordinate in $\tilde{X}$. That is, consider the function $\tilde{f} : \tilde{X} \to \mathbb{R}$, $y \mapsto f(x_1, ..., x_{j-1}, y, x_{j+1}, ..., x_n)$. Let $\tilde{\phi}$ be the restriction of ϕ to $Rg(\tilde{f})$ ($\subseteq Rg(f)$). We now replace the primitives X, f, ϕ and $\mathcal{F}$ with, respectively, $\tilde{X}$, $\tilde{f}$, $\tilde{\phi}$ and $\tilde{\mathcal{F}} = \{s : \tilde{X} \to \mathbb{R} : s \text{ is smooth } \& \ s'(x) \neq 0 \text{ for all } x \in \tilde{X}\}$. Note that we indeed have $\tilde{f} \in \tilde{\mathcal{F}}$ (show this using that $f \in \mathcal{F}$) and $\tilde{\phi} \circ \tilde{f} \in \tilde{\mathcal{F}}$ (show this using that $\phi \circ f \in \mathcal{F}$). As $\tilde{X}$ is one-dimensional, Claim 2 applies to these modified primitives. So, $\tilde{\phi}$ is smooth with $\tilde{\phi}' > 0$. Thus, as ϕ coincides with $\tilde{\phi}$ on $Rg(\tilde{f})$, ϕ is smooth with $\phi' > 0$ on $Rg(\tilde{f})$, hence at t.

2. We now show all three conditions on $\mathcal{N}$:

- Condition (i) holds by definition of $\mathcal{N}$.
- To show (iii), fix an $f \in \mathcal{N}$ and a smooth $\phi : Rg(f) \to \mathbb{R}$ with $\phi(0) = 0$. If $\phi'(0) = 1$, then $\phi \circ f \in \mathcal{N}$, since $\phi \circ f(\overline{x}) = \phi(0) = 0$, and since $(\phi \circ f)'(\overline{x})$ exists (as $f'(\overline{x})$ and ϕ' exist) with $\|(\phi \circ f)'(\overline{x})\| = \|\phi'(f(\overline{x}))f'(\overline{x})\| = |\phi'(f(\overline{x}))| \, \|f'(\overline{x})\| = 1 \times 1 = 1$. If instead $\phi'(0) \neq 1$, then $\phi \circ f \notin \mathcal{N}$, because $\|(\phi \circ f)'(\overline{x})\| \neq 1$.
- To show (iii), fix an $f \in \mathcal{F}$. The increasing affine transformation $g = \frac{1}{\|f'(\overline{x})\|}(f - f(\overline{x}))$ belongs to $\mathcal{N}$, since $g(\overline{x}) = 0$, and $g'(\overline{x})$ exists (as $g'(\overline{x})$ exists) with $\|g'(\overline{x})\| = \frac{1}{\|f'(\overline{x})\|} \, \|f'(\overline{x})\| = 1$. ■

A.2 Proof of results in Section 2

Depending on one's taste, one can read all following proofs either with the main text's concrete setup in mind or with the generalised setup in mind – except of course for the few results about the classic risk attitude, which hold for the concrete setup only.

Proof of Proposition 1. Fix a prospect order $\succeq$ and a well-behaved welfare function W. First, if W VNM represents $\succeq$, i.e., if $\succeq$ ranks prospects by expected welfare, then Intrinsic Risk Neutrality holds obviously. Conversely, assume Intrinsic Risk Neutrality. We must show that W VNM represents $\succeq$. We fix $p, q \in \mathcal{P}$ and will prove that $p \succeq q \Leftrightarrow \mathbb{E}_p(W) \geq \mathbb{E}_q(W)$. As W is regular, $Rg(W)$ is an interval. So, $\mathbb{E}_p(W), \mathbb{E}_q(W) \in Rg(W)$, and thus there exist $x, y \in X$ such that $W(x) = \mathbb{E}_p(W)$ and $W(y) = \mathbb{E}_q(W)$. By Intrinsic Risk Neutrality, $x \sim p$ and $y \sim q$. So, $p \succeq q$ is equivalent to $x \succeq y$, hence to $W(x) \geq W(x)$ (as W is compatible with riskless comparisons), i.e., to $\mathbb{E}_p(W) \geq \mathbb{E}_q(W)$, as desired. ■

Some notation and lemmas will prepare our next proofs.

Notation. Given a welfare function W, let $\mathcal{P}^W$ be the set of *welfare prospects*, i.e., finite-support lotteries over $Rg(W)$ rather than X. To each prospect $p \in \mathcal{P}$ corresponds a welfare prospect in $\mathcal{P}^W$, denoted p^W, where for each $w \in W$ we define $p^W(w)$ as $p(W = w)$ $(= p(\{x \in X : W(x) = w\}))$.

Lemma 2 *Assume $\succeq$ has a VNM representation U, and $W : X \to \mathbb{R}$ is ordinally equivalent to U. Then:*

(a) For all prospects $p, q \in \mathcal{P}$, $p^W = q^W \Rightarrow p \sim q$.

(b) In particular, we can define an order $\succeq^W$ on $\mathcal{P}^W$ by letting $a \succeq^W b$ if and only if $p \succeq q$ for some (hence by (a) any) $p, q \in \mathcal{P}$ such that $p^W = a$ and $q^W = b$.

(c) $\succeq^W$ has a VNM representation given by the (strictly increasing) function $\phi : Rg(W) \to \mathbb{R}$ such that $U = \phi \circ W$.

(d) $\succeq^W$ satisfies CARA if and only if $\succeq$ and W satisfy CIRA.

(e) In particular, if CIRA holds, ϕ is linear or strictly concave or strictly convex.

Proof. Let $\succeq$, U and W be as assumed.

(a) Given the assumptions, the argument is (informally) that if p and q have the same welfare prospect (i.e., $p^W = q^W$), then they have the same 'utility prospect' (as utility is a one-to-one function of welfare), and hence the same expected utility, implying that $p \sim q$. Q.e.d.

(b) The order $\succeq^W$ is well-defined, as its definition does not depend on the choice of p and q by (a). Q.e.d.

(c) Let ϕ be as specified. For all $p \in \mathcal{P}$, we have $\mathbb{E}_p(U) = \mathbb{E}_{p^W}(\phi)$, since

$$\begin{aligned}\mathbb{E}_p(U) &= \sum_{x \in X} p(x)U(x) = \sum_{w \in \mathbb{R}} \sum_{x \in X : W(x) = w} p(x)U(x) \\ &= \sum_{w \in \mathbb{R}} \left(\sum_{x \in X : W(x) = w} p(x) \right) \phi(w) = \sum_{w \in \mathbb{R}} p^W(w)\phi(w) = \mathbb{E}_{p^W}(\phi).\end{aligned}$$

The claim now follows from the observation that, for any p^W and q^W in $\mathcal{P}^W$ (where $p, q \in \mathcal{P}$), $p^W \succeq^W q^W$ is equivalent to $p \succeq q$, hence to $\mathbb{E}_p(U) \geq \mathbb{E}_q(U)$, which reduces to $\mathbb{E}_{p^W}(\phi) \geq \mathbb{E}_{q^W}(\phi)$. Q.e.d..

(d) First assume $\succeq^W$ satisfies CARA. To show CIRA, consider any $\Delta > 0$, any $p, p' \in \mathcal{P}$, and any $\rho, \rho' \in X$, such that $p \sim \rho$, $p' \sim \rho'$, and $p(W = w) = p'(W = w + \Delta)$ for each $w \in \mathbb{R}$. Then $p^W \sim^W \rho^W$, $p'^W \sim^W \rho'^W$, and $p^W(w) = p'^W(w + \Delta)$. So, as $\succeq^W$ satisfies CARA, $\rho'^W = \rho^W + \Delta$, i.e., $W(\rho') = W(\rho) + \Delta$. This establishes CIRA.

Conversely, assume CIRA. Consider any $\Delta > 0$, $a, a' \in \mathcal{P}^W$, and $t, t' \in Rg(W)$ such that $a \sim^W t$, $a' \sim^W t'$, and $a(w) = a'(w + \Delta)$ for all $w \in \mathbb{R}$ (where $a(w)$ stands for 0 if $w \notin Rg(W)$ and $a'(w + \Delta)$ stands for 0 if $w + \Delta \notin Rg(W)$). Pick $p, p' \in \mathcal{P}$ and $\rho, \rho' \in X$ such that $p^W = a$, $p'^W = a'$, $W(\rho) = t$ and $W(\rho') = t'$. Then $p \sim \rho$, $p' \sim \rho'$, and $p(W = w) = p'(W = w + \Delta)$ for each $w \in Rg(W)$. So, by CIRA, $W(\rho') = W(\rho) + \Delta$, i.e., $t' = t + \Delta$. This shows that $\succeq^W$ satisfies CARA. Q.e.d.

(e) Assume CIRA. The property established in (d) can be shown to imply that the risk premium has the same sign for all non-certain prospects, i.e., is always zero or always positive or always negative. This easily implies that U is linear or strictly concave or strictly convex, respectively. ■

The next lemma is a well-known building block of the classical theory of risk aversion after Arrow (1965) and Pratt (1964), and will later be applied to the order $\succeq^W$ in Lemma 2.

Lemma 3 *If an order on the set of finite-support lotteries over a given real interval has a smooth VNM representation with everywhere positive derivative, then it satisfies CARA if and only if it has a VNM representation given by* $w \mapsto \frac{1}{\rho}(e^{\rho w} - 1)$ *for some* $\rho \in \mathbb{R}$.

If $\rho = 0$, then '$\frac{1}{\rho}(e^{\rho w} - 1)$' of course stands for w $(= \lim_{\rho \to 0} \frac{1}{\rho}(e^{\rho w} - 1))$. Although this lemma is well-known, we sketch the argument for completeness.

Proof. Consider an order $\succeq^*$ on the set $\mathcal{P}^*$ of finite-support lotteries over a given interval $I \subseteq \mathbb{R}$, with a smooth VNM representation ϕ. For each $\rho \in \mathbb{R}$ let $\phi_\rho : I \to \mathbb{R}$ be the function $w \mapsto \frac{1}{\rho}(e^{\rho w} - 1)$. The proof goes in two steps.

Claim 1: $\succeq^*$ satisfies CARA if and only if there exists a $\rho \in \mathbb{R}$ such that ϕ solves the differential equation '$f'' = \rho f'$' on I, the solutions of which are the affine transformations of ϕ_ρ.

By the fundamental result of Arrow (1965) and Pratt (1964), $\succeq^*$ satisfies CARA if and only if the function ϕ''/ϕ' is constant, which implies the 'if and only if' claim. The set of solutions to the differential equation '$f'' = \rho f'$' (on I) is well-known:

- If $\rho \neq 0$, then the solutions are the affine transformations of the function $w \mapsto e^{\rho w}$.
- If $\rho = 0$, so that "$f'' = \rho f'$' reduces to '$f'' = 0$', then the solutions are the affine transformations of the function $w \mapsto w$.

So, whether $\rho \neq$ or $\rho = 0$, the solutions are the affine transformations of ϕ_ρ. Q.e.d..

Claim 2: If ϕ' is everywhere positive, then $\succeq^*$ satisfies CARA if and only if there exists a $\rho \in \mathbb{R}$ such that ϕ_ρ VNM represents $\succeq^*$.

Assume ϕ' is everywhere positive. Then ϕ and ϕ_ρ are two increasing functions, hence are increasing transformations of one another. By Claim 1, $\succeq^*$ satisfies CARA if and only if there exists a $\rho \in \mathbb{R}$ such that ϕ_ρ and ϕ are (now increasing) affine transformations of one another, or equivalently such that ϕ_ρ (like ϕ) VNM represents $\succeq^*$. ■

Proof of Proposition 2. Consider any regular $\succeq$ and well-behaved W.

1. In this part we assume that $W = \log(\rho U + 1)/\rho$ for a VNM representation U of $\succeq$ and a $\rho \in \mathbb{R}$ such that $\rho U + 1 > 0$, and we prove that W satisfies CIRA. Note first that $U = (e^{\rho W} - 1)/\rho$. Thus, $U = \phi_\rho \circ W$, where ϕ_ρ is the function on $Rg(W)$ given by $w \mapsto (e^{\rho w} - 1)/\rho$. Let $\succeq^W$ be the order over welfare prospects defined in Lemma 2. By Lemma 2(c), ϕ_ρ VNM represents $\succeq^W$. So $\succeq^W$ satisfies CARA by Lemma 3. This implies CIRA by Lemma 2(d). Q.e.d.

2. Conversely, assume CIRA. We show the existence of a VNM representation U of $\succeq$ and a $\rho \in \mathbb{R}$ such that $\rho U + 1 > 0$ and $W = \log(\rho U + 1)/\rho$. Define $\succeq^W$ and the transformations $\phi_\rho : Rg(W) \to \mathbb{R}$ ($\rho \in \mathbb{R}$) as before. Being regular, $\succeq$ has a regular VNM representation $\tilde{U} : X \to \mathbb{R}$. As $\tilde{U}$ and W are regular and ordinally equivalent, $\tilde{U} = \phi \circ W$ for a smooth transformation $\phi : Rg(W) \to \mathbb{R}$ with $\phi' > 0$. ϕ VNM represents $\succeq^W$ by Lemma 2(c). CIRA implies that $\succeq^W$ satisfies CARA by Lemma 2(d). Hence, by Lemma 3, there exists a $\rho \in \mathbb{R}$ such that ϕ_ρ VNM represents $\succeq^W$. As ϕ_ρ and ϕ both VNM represent $\succeq^W$, ϕ_ρ is an increasing affine transformation of ϕ. So, the function $U := \phi_\rho \circ W$ is an increasing affine transformation of $\tilde{U}$ ($= \phi \circ W$). Hence, not only $\tilde{U}$ but also U VNM represents $\succeq$. We have $\rho U + 1 > 0$, as $\rho U + 1 = \rho(\phi_\rho \circ W) + 1 > \rho(-1/\rho) + 1 = 0$. Finally, $W = \phi_\rho^{-1} \circ U = \log(\rho U + 1)/\rho$. ■

We now formally restate and prove Fact 2, which claims that CIRA can explain violations of CARA whenever the welfare function is not of a special type. More precisely:

Fact 2 *(formal statement): For any regular prospect order $\succeq$ with $X \subseteq \mathbb{R}$, CARA is violated if CIRA holds for a well-behaved welfare function W that is neither linear nor exponential nor logarithmic nor a logarithmic function of an exponential function.*

Fact 2 follows from a more general characterisation:

Proposition 8 *Given a regular prospect order $\succeq$ and a well-behaved welfare function W satisfying CIRA, where $X \subseteq \mathbb{R}$, CARA holds <u>if and only if</u> W is (i) a linear or exponential function with $\rho_W = 0$, or (ii) the base-e^{ρ_W} logarithm of such a function with $\rho_W \neq 0$.*

For instance, if $\rho_W = -1$ (a form of intrinsic risk aversion), then CARA holds precisely if welfare takes the form $W = \frac{\log(V)}{\log(e^{-1})} = -\log(V)$ for some linear or exponential function V.

The proof of Proposition 8 will draw on the following well-known result, which is an obvious variant of Lemma 3 (without 'positive derivative' restriction):

Lemma 4 *If an order on the set of finite-support lotteries over a given real interval has a smooth VNM representation U, then it satisfies CARA if and only if U is linear or exponential.*

Proof of Proposition 8. Assume $X \subseteq \mathbb{R}$, $\succeq$ is regular, W is well-behaved, and CIRA holds. By CIRA, ρ_W is constant (see Corollary 1).

1. First, assume CARA. We show that W takes one of the special forms. CIRA implies that there is a VNM representation U of $\succeq$ such that $W = \log(\rho_W U + 1)/\rho_W$, interpreted as $W = U$ if $\rho_W = 0$ (Proposition 2 and Corollary 1). Meanwhile CARA implies that U is linear or exponential (Lemma 4). So, if $\rho_W = 0$, then W $(= U)$ is linear or exponential, while if $\rho_W \neq 0$, then

$$W = \log(\rho_W U + 1)/\rho_W = \log(\rho_W U + 1)/\log(e^{\rho_W}),$$

i.e., W is the base-e^{ρ_W} logarithm of a function that is linear (if U is linear) or exponential (if U is exponential). Thus W takes one of the special forms.

2. Now assume W takes one of these forms. By CIRA, $\succeq$ is VNM representable by a function U that we can scale such that $U = W$ if $\rho_W = 0$ and $U = ke^{\rho_W W}$ for a $k \in \mathbb{R}$ of same sign as ρ_W if $\rho_W \neq 0$ (see non-fully 2). First assume $\rho_W = 0$. Then $U = W$, and W is linear or exponential. So U is linear or exponential, implying CARA (Lemma 4).

Now assume $\rho_W \neq 0$. Then $U = ke^{\rho_W W}$, and $W = \frac{1}{\rho_W} \log V$ for a linear or exponential function V. Thus

$$U = e^{\rho_W W} = e^{\rho_W \frac{1}{\rho_W} \log V} = e^{\log V} = V.$$

So U is linear or exponential, again implying CARA (Lemma 4). ■

A.3 Proof of Proposition 3

Assume $X \subseteq \mathbb{R}$. Let $\succeq$ be regular and W well-behaved. As $\succeq$ is regular, it has a regular VNM representation U. As W and U are ordinally equivalent and regular, the (unique) function $\phi : Rg(W) \to \mathbb{R}$ such that $U = \phi(W)$ is smooth with $\phi' > 0$. Differentiation yields

$$U' = \phi'(W)W' \text{ and } U'' = \phi''(W)W'^2 + \phi'(W)W''.$$

Hence the classical risk attitude $\rho_{AP} = \frac{U''}{U'}$ is given by

$$\rho_{AP} = \frac{\phi'(W)W'' + \phi''(W)W'^2}{\phi'(W)W'} = \frac{W''}{W'} + \frac{\phi''(W)}{\phi'(W)}W' = \frac{W''}{W'} + \rho_W W'. \blacksquare$$

A.4 Proof of the results in Section 4

The following lemma implies the 'Remark' in Section 4:

Lemma 5 *A prospect order $\succeq$ with a VNM representation U is broad-ranging if and only if U is unbounded, i.e., satisfies $\sup U = \infty$ or $\inf U = -\infty$.*

Proof. Assume a prospect order $\succeq$ has a VNM representation U.

1. First let U be unbounded. Without loss of generality, suppose $\sup U = \infty$ (an analogous proof works if instead $\inf U = -\infty$). To prove that $\succeq$ is broad-ranging, consider situations $x, y \in X$ with $x \succeq y$. So $U(x) \geq U(y)$. As $\sup U = \infty$, there is a situation $z \in X$ such that $U(z) - U(x) > U(x) - U(y)$. It easily follows that $\frac{1}{2}U(z) + \frac{1}{2}U(y) > U(x)$. So, as U VNM represents $\succeq$, $z_{\frac{1}{2}}y_{\frac{1}{2}} \succ x$.

2. Conversely, let $\succeq$ be broad-ranging. In particular, not all situations in X are equally good. Pick any $x \succ y$ in X, and write $\Delta = U(x) - U(y)$ (> 0). For $j = 0, 1, ...$ define situations x_j and y_j with $U(x_j) - U(y_j) \geq 2^j\Delta$ recursively as follows. First, $x_0 = x$ and $y_0 = y$. Clearly $U(\overline{x}) - U(y_0) \geq 2^0\Delta$ (in fact, '$\geq$' could be replaced by '='). Now consider $j \geq 0$ and suppose x_j and y_j are defined, with $U(x_j) - U(y_j) \geq 2^j\Delta$. As $\succeq$ is broad-ranging, there exists a $g \in X$ such that $g_{\frac{1}{2}}y_{\frac{1}{2}} \succ x$ ('case 1') or there exists a $b \in X$ such that $y \succ b_{\frac{1}{2}}x_{\frac{1}{2}}$ ('case 2').

First assume case 1. Put $x_{j+1} = g$ and $y_{j+1} = y_j$. So, $\frac{1}{2}U(x_{j+1}) + \frac{1}{2}U(y_{j+1}) > U(x_j)$, and thus

$$\frac{1}{2}U(x_{j+1}) - \frac{1}{2}U(y_{j+1}) > U(x_j) - U(y_{j+1}) = U(x_j) - U(y_j) \geq 2^j\Delta.$$

Hence $U(x_{j+1}) - U(y_{j+1}) \geq 2^{j+1}\Delta$, as desired.

Now assume case 2 but not case 1. Put $x_{j+1} = x_j$ and $y_{j+1} = b$. So, $\frac{1}{2}U(x_{j+1}) + \frac{1}{2}U(y_{j+1}) < U(y_j)$, and thus

$$\frac{1}{2}U(y_{j+1}) - \frac{1}{2}U(x_{j+1}) < U(y_j) - U(x_{j+1}) = U(y_j) - U(x_j) \leq 2^j\Delta.$$

Hence again $U(x_{j+1}) - U(y_{j+1}) \geq 2^{j+1}\Delta$, as desired.

As $j \to \infty$, we have $2^j\Delta \to \infty$, and so $U(x_j) - U(y_j) \to \infty$. So, $\sup U = \infty$ or $\inf U = -\infty$. ■

We now prove Proposition 4 about welfare measurement under CIRA and Full-range, and Theorem 1 about welfare measurement under CIRA, Full-range and Normalisation. In fact, to complete the picture, we shall also prove a third result, about welfare measurement under CIRA and Normalisation:

Proposition 9 *Given a regular prospect order $\succeq$, a well-behaved welfare function W satisfies CIRA and Normalisation if and only if it takes the form*

$$W = \log(\rho U + 1)/\rho$$

for the normalised VNM representation U of $\succeq$ and some 'intrinsic risk attitude' $\rho \in \mathbb{R}$ such that $\rho U + 1 > 0$.

We will derive these three results from Proposition 2 via the following key insight:

Lemma 6 *For any regular prospect order $\succeq$ and any welfare function of the form $W = \log(\rho U + 1)/\rho$ for a VNM representation U of $\succeq$ and a $\rho \in \mathbb{R}$ such that $\rho U + 1 > 0$,*

(a) W satisfies Full-range if and only if $0 \in Rg(U)$ and

$$\rho = \begin{cases} \frac{-1}{\sup U} \ (< 0) & \text{if } \sup U \neq \infty \\ \frac{-1}{\inf U} \ (> 0) & \text{if } \inf U \neq -\infty \\ 0 & \text{if } \sup U = \infty \text{ and } \inf U = -\infty, \end{cases}$$

assuming $\succeq$ is broad-ranging,

(b) W satisfies Normalisation if and only if U is normalised,

(c) W satisfies Full-range and Normalisation if and only if U is normalised and ρ is given as in (b), assuming $\succeq$ is broad-ranging.

Proof. Let $\succeq$, U and ρ be as specified.

(a) Assume $\succeq$ is broad-ranging. Since U is regular, $Rg(U)$ is an open interval (a, b), where $a = \inf U$ and $b = \sup U$. As $\succeq$ is broad-ranging, U is unbounded (by Lemma 5), so that either $a = -\infty$ or $b = \infty$ (possibly both).

If it is *not* the case that $a < 0 < b$, then $0 \notin Rg(U)$, and thus $0 \notin Rg(W)$; hence both sides of the claimed equivalence are violated, and thus the equivalence holds.

Now assume that $a < 0 < b$. As $Rg(U) = (a, b)$ and $W = \log(\rho U + 1)/\rho$, $Rg(W)$ is the open interval with the boundaries

$$\inf W = \lim_{u \downarrow a} \log(\rho u + 1)/\rho \text{ and } \sup W = \lim_{u \uparrow b} \log(\rho u + 1)/\rho.$$

This uses that $\log(\rho u + 1)/\rho$ is a smooth and strictly increasing function of u, whether ρ is negative, positive, or zero (if $\rho = 0$ then $\log(\rho u + 1)/\rho$ stands for u, as usual). Since Full-range means that $Rg(W) = \mathbb{R}$,

$$\text{Full-range} \ \Leftrightarrow \ [\lim_{u \downarrow a} \log(\rho u + 1)/\rho = -\infty \ \& \ \lim_{u \uparrow b} \log(\rho u + 1)/\rho = \infty].$$

If ρ is positive, we must show that Full-range is equivalent to $\rho = -\frac{1}{a}$. This holds because

$$\begin{aligned} \text{Full-range} \ &\Leftrightarrow \ \lim_{u \downarrow a} \log(\rho u + 1) = -\infty \text{ and } \lim_{u \uparrow b} \log(\rho u + 1) = \infty \\ &\Leftrightarrow \ \rho a + 1 = 0 \text{ and } \rho b + 1 = \infty \\ &\Leftrightarrow \ \rho = -\tfrac{1}{a} \text{ and } b = \infty \\ &\Leftrightarrow \ \rho = -\tfrac{1}{a}, \end{aligned}$$

where we could drop 'and $b = \infty$' since this is implied by a's finiteness and U's unboundedness. Analogously, if ρ is negative, we show that Full-range is equivalent to $\rho = -\frac{1}{b}$:

$$\begin{aligned} \text{Full-range} \ &\Leftrightarrow \ \lim_{u \downarrow a} \log(\rho u + 1) = \infty \text{ and } \lim_{u \uparrow b} \log(\rho u + 1) = -\infty \\ &\Leftrightarrow \ \rho a + 1 = \infty \text{ and } \rho b + 1 = 0 \\ &\Leftrightarrow \ a = -\infty \text{ and } \rho = -\tfrac{1}{b} \\ &\Leftrightarrow \ \rho = -\tfrac{1}{b}. \end{aligned}$$

Finally, if $\rho = 0$, we must show that Full-range is equivalent to $\rho = 0$. This holds because both sides of the equivalence are true: $\rho = 0$ by assumption, and Full-range holds since $W = U$, so that $Rg(W) = Rg(U) = \mathbb{R}$.

(b) We must show that W is normalised if only if U is normalised. This follows from two observations. First, as $W = \frac{1}{\rho}\log(\rho U + 1)$, W takes the value 0 exactly where U takes the value 0. Second,

$$W' = \frac{1}{\rho}\log'(\rho U + 1)(\rho U + 1)' = \frac{1}{\rho(\rho U + 1)}\rho U' = \frac{1}{\rho U + 1}U',$$

so that W' and U' coincide wherever W (or equivalently U) is 0.

(c) Assume $\succeq$ is broad-ranging. By parts (a) and (b), Full-range and Normalisation jointly bold if and only if U is normalised, $0 \in Rg(U)$, and ρ is given as in (b). In this equivalence, we can drop the condition that $0 \in Rg(U)$, which follows from the normalisation of U. ■

Proof of Proposition 4. This result follows from Proposition 2 and Lemma 6(a). Let $\succeq$ be regular and broad-ranging, and W well-behaved. By Proposition 2, W satisfies CIRA if and only if $W = \log(\rho U + 1)/\rho$ for some VNM representation U of $\succeq$ such that $\rho U + 1 > 0$. By Lemma 6(a), this equivalence continues to hold if on the left we add Full-range, and on the right we add that $0 \in Rg(U)$ and that ρ takes the specified value, while dropping the redundant condition $\rho U + 1 > 0$ (which follows from the value of ρ). This yields the desired equivalence. ■

Proof of Proposition 9. This result follows from Proposition 2 and Lemma 6(b). Let $\succeq$ be regular, and W well-behaved. By Proposition 2, W satisfies CIRA if and only if $W = \log(\rho U + 1)/\rho$ for some VNM representation U of $\succeq$ such that $\rho U + 1 > 0$. By Lemma 6(b), this equivalence continues to hold if on the left we add Normalisation, and on the right we add that U is normalised. This yields the intended equivalence. ■

Proof of Theorem 1 (in both versions). This result follows from Proposition 2 and Lemma 6(c). Let $\succeq$ be regular and broad-ranging, and W well-behaved. By Proposition 2, W satisfies CIRA if and only if $W = \log(\rho U + 1)/\rho$ for some VNM representation U of $\succeq$ such that $\rho U + 1 > 0$. By Lemma 6(c), this equivalence continues to hold if on the left we add Full-range and Normalisation, and on the right we add that $0 \in Rg(U)$, that ρ takes the specified value, and that U is normalised, while dropping the redundant conditions that $0 \in Rg(U)$ (which follows from the normalisation of U) and that $\rho U + 1 > 0$ (which follows from the value of ρ). This yields the theorem in its initial version. This version immediately implies the theorem's second version, because the welfare function $W = \log(\rho U + 1)/\rho$ characterised in the initial version is (by its uniqueness) the *revealed* function $W_{\succeq}$, and because the revealed intrinsic risk attitude $\rho_{\succeq}$ coincides with ρ (by Corollary **??**). ■

A.5 Proof of Facts 4 and 5

The proof of Fact 4 is prepared by two simple observations:

Lemma 7 *If a prospect order $\succeq$ is regular, then each prospect $p \in \mathcal{P}$ has a certainty equivalent, i.e., an $x_p \in X$ such that $x_p \sim p$.*

Proof. Suppose $\succeq$ is regular and $p \in \mathcal{P}$. Since $\succeq$ is regular, it has a is regular VNM representation U. As U is regular, $Rg(U)$ is an interval, and so $\mathbb{E}_p(U) \in Rg(U)$. So, there is an $x \in X$ such that $\mathbb{E}_p(U) = U(x)$, and thus $p \sim x$. ■

Lemma 8 *Given regular individual prospect orders $\succeq_1, \ldots, \succeq_n$ such that diversity holds, and a welfare function under risk $\mathbf{W}$ satisfying the Pareto Principle, each prospect $p \in \mathcal{P}$ has a social certainty equivalent, i.e., an $x_p \in X$ such that $\mathbf{W}(x_p) = \mathbf{W}(p)$.*

Proof. Suppose the assumptions hold. Fix a prospect $p \in \mathcal{P}$. For each i, since $\succeq_i$ is regular, x has a certainty equivalent for person i, i.e., an $x_i \in X$ such that $p \sim_i x_i$ (by Lemma 7). By diversity, there is an $x \in X$ such that $x_i \sim_i x$ for all i. Thus $p \sim_i x$ for all i. So, by the Pareto Principle, $\mathbf{W}(p) = \mathbf{W}(x)$, i.e., x is a social certainty equivalent of p. ■

Proof of Fact 4. As assumed in Section 5, let $W_1 \ldots, W_n$ be the welfare functions revealed by (regular and broad-ranging) prospect orders $\succeq_1, \ldots, \succeq_n$.

(a) If $\mathbf{W} = \sum_i W_i^{\text{equ}}$, then the Pareto Principle holds, because each W_i^{equ} is increasing in $\succeq_i$.

(b) If $\mathbf{W} = \sum_i W_i^{\text{exp}} = (\sum_i W_i)^{\text{exp}}$, then $\mathbf{W}$ is the expectational extension of the riskfree social welfare function $\mathbf{W}|_X$ $(= \sum_i W_i)$, and thus it is VNM-rational as well as intrinsic risk neutral, i.e., neutral to risk in social welfare.

(c) First, assume $\mathbf{W} = \sum_i U_i^{\text{exp}}$ $(= (\sum_i U_i)^{\text{exp}} = \sum_i U_i^{\text{equ}})$ for VNM representations $U_1, \ldots, U_n$ of $\succeq_1, \ldots, \succeq_n$. Since $\mathbf{W} = \sum_i U_i^{\text{equ}}$ and each U_i^{equ} is increasing in $\succeq_i$, the Pareto Principle holds, just as in (a). Further, since $\mathbf{W} = (\sum_i U_i)^{\text{exp}}$, $\mathbf{W}$ is the expectational extension of the riskfree social welfare function $\mathbf{W}|_X$ $(= \sum_i U_i)$, hence is VNM-rational and intrinsic risk neutral, again like in (b).

Conversely, suppose $\mathbf{W}$ satisfies the three conditions, and assume diversity. Consider the social welfare *order* under risk, $\succeq_{\mathbf{W}}$. It satisfies the VNM-axioms, and it is Paretian w.r.t. $\succeq_1, ..., \succeq_n$. So, by Harsanyi's Theorem (1955), there exist VNM representations $U_1, \ldots, U_n$ of $\succeq_1, \ldots, \succeq_n$ such that the function $\sum_i U_i$ VNM represents $\succeq_{\mathbf{W}}$. Meanwhile, by Social Risk Neutrality, $\succeq_{\mathbf{W}}$ is also VNM representable by the riskless social welfare function $\mathbf{W}|_X$ (see Proposition 1). Since $\sum_i U_i$ and $\mathbf{W}|_X$ both VNM represent the same prospect order $\succeq_{\mathbf{W}}$, there are $a > 0$ and $b \in \mathbb{R}$ such that $\mathbf{W}|_X = a \sum_i U_i + b$. For each i, since U_i VNM represents $\succeq_i$, so does $\tilde{U}_i = aU_i + \frac{1}{n}b$. It remains to show that $\mathbf{W} = \sum_i \tilde{U}_i^{\text{exp}}$. Write $\mathbf{W}^* = \sum_i \tilde{U}_i^{\text{exp}}$.

To show that $\mathbf{W} = \mathbf{W}^*$, note first that $\mathbf{W}|_X = \mathbf{W}^*|_X$, since $\mathbf{W}|_X$ and $\mathbf{W}^*|_X$ each coincide with $\sum_i \tilde{U}_i$. More generally, fix a $p \in \mathcal{P}$, and let us show that $\mathbf{W}(p) = \mathbf{W}^*(p)$.

Using diversity and the Pareto Principle, p has a social certainty equivalent x_p (Lemma 8). Since $p \sim_{\mathbf{W}} x_p$ and since $\mathbf{W}$ and $\mathbf{W}^*$ both represent $\succeq_{\mathbf{W}}$, we have $\mathbf{W}(p) = \mathbf{W}(x_p)$ and $\mathbf{W}^*(p) = \mathbf{W}^*(x_p)$. Meanwhile, $\mathbf{W}(x_p) = \mathbf{W}^*(x_p)$, since $\mathbf{W}|_X = \mathbf{W}^*|_X$. So, $\mathbf{W}(p) = \mathbf{W}^*(p)$. ■

Proof of Fact 5. Risk-impartial Utilitarianism is socially rational because $\succeq_{\mathbf{W}}$ is a VNM order (with VNM utility function given by $U = (e^{\rho W} - 1)/\rho$ or $W = \log(\rho U + 1)/\rho$, where $W = \sum_i W_i$ and $\rho = \frac{1}{n}\sum_i \rho_i$). Social CIRA holds because $\mathbf{W}$ has a constant intrinsic risk attitude (given by ρ). ■

A.6 Proof of results in Section 6

To prove Proposition 5, we first present a simple characterisation of admissible welfare functions under the hypotheses of CIRA and Full-range, which corresponds to the characterisation presented under case 5 in the 'Summary' in Section 6:[49]

Lemma 9 *Given a regular prospect order $\succeq$, a well-behaved welfare function W satisfies CIRA and Full-range if and only if*

- *either $W = U$ for some VNM representation U of $\succeq$ with full range $\mathbb{R}$ (case of intrinsic risk neutrality)*
- *or $W = \sigma \log U$ for some VNM representation U of $\succeq$ with range $(0, \infty)$ and some $\sigma > 0$ (case of intrinsic risk proneness)*
- *or $W = -\sigma \log(-U)$ for some VNM representation U of $\succeq$ with range $(-\infty, 0)$ and some $\sigma > 0$ (case of intrinsic risk neutrality).*

Proof. As long as $\succeq$ is broad-ranging, this claim can be reduced to that in case 5 of the 'Summary' in Section 6, through the following simple reparametrisation:

- If $\rho_U = 0$, then set $V = U$. Note that $Rg(U) = \mathbb{R}$, so that $Rg(W) = \mathbb{R}$.
- If $\rho_U > 0$, then set $V = \rho_U U + 1$ and $\sigma = 1/\rho_U$. Note that U is bounded below and that $\rho_U = -1/\inf U$, so that $\inf V = 0$.
- If $\rho_U < 0$, then set $V = -\rho_U U - 1$ and $\sigma = -1/\rho_U$. Note that U is bounded above and that $\rho_U = -1/\sup U$, so that $\sup V = 0$.

If $\succeq$ is *not* broad-ranging, the claim holds by the argument in footnote 49. ■

Proof of Proposition 5. Let $\succeq$ be regular and broad-ranging, and $\mathbb{W}$ be the set of well-behaved welfare functions satisfying CIRA and Full-range. Note that $\mathbb{W} \neq \varnothing$ – this for

[49] Lemma 9 is in fact more general than the statement in case 5 of the 'Summary', since it allows $\succeq$ not to be broad-ranging, in which case the equivalence in Lemma 9 still holds, because both sides of the equivalence are then false. Why are they both false? If $\succeq$ is not brad-ranging, then U has a bounded range by Lemma 5. So, the right-hand-side of the equivalence is obviously false. And so is the left-hand-side, because CIRA implies that $W = \log(\rho U + 1)/\rho$ for some $\rho \in \mathbb{R}$ (see Proposition 2), which by the boundedness of U's range implies that W has a bounded range, i.e., violates Full-range.

instance follows from case 5 in the 'Summary' in Section 6. So we may pick a $W \in \mathbb{W}$. We must show that $\mathbb{W} = \{aW + b : a > 0, b \in \mathbb{R}\}$. This is done by showing both set inclusions in turn.

Claim 1: $\mathbb{W} \supseteq \{aW + b : a > 0, b \in \mathbb{R}\}$.

We fix $a > 0$ and $b \in \mathbb{R}$, and show that $aW+b \in \mathbb{W}$. Since $W \in \mathbb{W}$, W is well-behaved, satisfies CIRA, and satisfies Full-range. So, $aW + b$ is also well-behaved, satisfies CIRA, and satisfies Full-range. Thus $aW + b \in \mathbb{W}$. Q.e.d.

Claim 2: $\mathbb{W} \subseteq \{aW + b : a > 0, b \in \mathbb{R}\}$.

We now fix a $\hat{W} \in \mathbb{W}$, and show that $\hat{W} \in \{aW + b : a > 0, b \in \mathbb{R}\}$. As $\succeq$ is broad-ranging, the VNM representations in $\mathbb{U}_{\succeq}$ are unbounded by Lemma 9. There are three cases.

Case 1: all functions in $\mathbb{U}_{\succeq}$ have range $\mathbb{R}$. Then, by Lemma 9, $W = U$ and $\hat{W} = \hat{U}$ for VNM representations U and $\hat{U}$ of $\succeq$ (with range $\mathbb{R}$). Thus $\hat{W}$ is an increasing affine transformation of W, i.e., $\hat{W} \in \{aW + b : a > 0, b \in \mathbb{R}\}$.

Case 2: all functions in $\mathbb{U}_{\succeq}$ are bounded below and unbounded above. Then, by Lemma 9, $W = \sigma \log V$ and $\hat{W} = \hat{\sigma} \log \hat{V}$ for VNM representations V and $\hat{V}$ of $\succeq$ with range $(0, \infty)$ and $\sigma, \hat{\sigma} > 0$. Note that $\hat{V}$ is an increasing affine transformation of V with $\inf \hat{V} = \inf V = 0$. So, there is a $k > 0$ such that $\hat{V} = kV$. Meanwhile, since $W = \sigma \log V$, we have $V = e^{W/\sigma}$, and thus $\hat{V} = ke^{W/\sigma}$. Now

$$\hat{W} = \hat{\sigma} \log \hat{V} = \hat{\sigma} \log(ke^{W/\sigma}) = \hat{\sigma}(\log k + \log e^{W/\sigma}) = \hat{\sigma} \log k + \frac{\hat{\sigma}}{\sigma} W.$$

So, $\hat{W} \in \{aW + b : a > 0, b \in \mathbb{R}\}$.

Case 3: all functions in $\mathbb{U}_{\succeq}$ are bounded above and unbounded below. Then, by Lemma 9, $W = -\sigma \log(-V)$ and $\hat{W} = -\hat{\sigma} \log(-\hat{V})$ for VNM representations V and $\hat{V}$ of $\succeq$ with range $(-\infty, 0)$ and $\sigma, \hat{\sigma} > 0$. Note that $\hat{V}$ is an increasing affine transformation of V with $\sup \hat{V} = \sup V = 0$. So, there is a $k > 0$ such that $\hat{V} = kV$. As $W = -\sigma \log(-V)$, we have $V = -e^{-W/\sigma}$, and thus $\hat{V} = -ke^{-W/\sigma}$. Finally,

$$\hat{W} = -\hat{\sigma} \log(-\hat{V}) = -\hat{\sigma} \log(ke^{-W/\sigma}) = -\hat{\sigma}(\log k + \log e^{-W/\sigma}) = -\hat{\sigma} \log k + \frac{\hat{\sigma}}{\sigma} W.$$

So, $\hat{W} \in \{aW + b : a > 0, b \in \mathbb{R}\}$. ■

Proof of Proposition 6. While this result could be shown by manipulating the functional expression of welfare in case 3, a more basic argument focuses on the hypotheses CIRA and well-behavedness themselves: each of them is clearly preserved when replacing W with $\hat{W} = aW + b$ (for fixed $a > 0$ and $b \in \mathbb{R}$). ■

A.7 Proof of Proposition 7

Assume an order structure $(\succeq_z)$ satisfying (a) and (b), and let $\succeq \equiv \succeq_z$ be regular and W well-behaved. Let Δ, p, w_p, q, w_q be as in CIRA. Let $p(W = w) = q(W = w + \Delta)$ for all

$w \in \mathbb{R}$. We must show that $w_q = w_p + \Delta$. By (a), there exists an order $\succeq^*$ over welfare-change lotteries (i.e., finite-support lotteries on $\mathbb{R}$) such that $\tilde{p} \succeq_{\tilde{z}} \tilde{q} \Leftrightarrow \tilde{p}_{\Delta W|\tilde{z}} \succeq^* \tilde{q}_{\Delta W|\tilde{z}}$ for all $\tilde{p}, \tilde{q} \in \mathcal{P}$ and $\tilde{z} \in X$. Pick situations $z, z' \in X$ such that $W(z') = W(z) + \Delta$ (they exist by Full-range). Pick certainty equivalents $x_p, x_q \in X$ of p resp. q (they exist as $\succeq$ is regular). Now $p \sim x_p$, and so $p \sim_z x_p$, whence $p_{\Delta W|z} \sim^* W(x_p) - W(z) = w_p - W(x_0)$ (identifying the welfare change $W(x_p) - W(z) = w_p - W(x_0)$ with a riskless welfare-change lottery). Analogously, $q \sim x_q$, and so $q \sim_{z'} y$, whence $q_{\Delta W|z'} \sim^* W(x_q) - W(z') = w_q - W(z')$. Further, $p_{\Delta W|z} = q_{\Delta W|z'}$, since q's welfare prospect equals p's shifted by Δ, and z''s welfare equals z's shifted by Δ. In sum, $p_{\Delta W|z} \sim^* w_p - W(z)$, $q_{\Delta W|z'} \sim^* w_q - W(z')$, and $p_{\Delta W|z} = q_{\Delta W|z'}$. Thus $w_p - W(z) \sim^* w_q - W(z')$. So $w_p - W(z) = w_q - W(z')$, since indifferent welfare changes are identical (as $\succeq$ is regular and W is well-behaved). Thus, $w_q - w_p = W(z') - W(z) = \Delta$. ■

A.8 Proof of Theorem 2 via Theorem 1

The following lemma prepares the reduction of Theorem 2 to Theorem 1.

Lemma 10 *For any prospect order $\succeq$, any instance of the generalised conditions Full-range$_D$, Normalisation$_{r,s}$ and CIRA$_\tau$, any welfare function W such that $Rg(W) \subseteq D$ (ensuring that $\tau \circ W$ is defined), and any increasing affine transformation W^* of $\tau \circ W$,*

(a) W is well-behaved if and only if W^* is well-behaved,

(b) W satisfies Full-range$_D$ if and only if W^* satisfies Full-range,

(c) W satisfies Normalisation$_{r,s}$ if and only if W^* satisfies Normalisation, assuming $W^* = (\tau \circ W - \tau(r))/(s\tau'(r))$,

(d) W satisfies CIRA$_\tau$ if and only if W satisfies CIRA.

Proof. Consider D, r, s, τ,W and W^* as specified. Let $\phi : D \to \mathbb{R}$ be the increasing affine transformation of τ such that $W^* = \phi \circ W$. Since τ is a smooth and positively differentiable function from D onto $\mathbb{R}$, so is ϕ. By basic analysis, it follows that ϕ^{-1} exists (so that $W = \phi^{-1} \circ W^*$) and that ϕ^{-1} is a smooth and positively differentiable function from $\mathbb{R}$ onto D.

(a) Recall that well-behavedness is the conjunction of compatibility with riskless comparisons and regularity. So the claim follows from two facts:

- W^* is compatible with riskless comparisons if and only if W is so, since W and W^* are ordinally equivalent.
- W^* is regular if and only if W is regular, since W and W^* are smooth positively differentiable transformations of one another.

(b) We have to show that $Rg(W) = D \Leftrightarrow Rg(W^*) = \mathbb{R}$. Note that $Rg(W^*) = \mathbb{R} \Leftrightarrow Rg(\tau \circ W) = \mathbb{R}$, as W^* is an increasing affine transformation of $\tau \circ W$. So it suffices to show that $Rg(W) = D \Leftrightarrow Rg(\tau \circ W) = \mathbb{R}$. This equivalence holds because, firstly, if

$Rg(W) = D$ then $Rg(\tau \circ W) = \tau(Rg(W)) = \tau(D) = \mathbb{R}$, and secondly, if $Rg(W) \neq D$ then $Rg(\tau \circ W) = \tau(Rg(W)) \neq \tau(D) = \mathbb{R}$.

(c) Suppose $W^* = (\tau \circ W - \tau(r))/(s\tau'(r))$. In other words, $W^* = \phi \circ W$ where $\phi = (\tau(\cdot) - \tau(r))/(s\tau'(r))$. As a preliminary, consider the smooth transformation $\tilde{\phi}$ defined on $\tilde{D} = \{(d-r)/s : d \in D\}$ by $\tilde{\phi}(t) = \phi(st + r)$ for all $t \in \tilde{D}$. We have $\tilde{\phi}(0) = 0$ and $\tilde{\phi}'(0) = 1$, since $\tilde{\phi}(0) = \phi(r) = 0$ and $\tilde{\phi}'(t) = \frac{s\tau'(ts+r)}{s\tau'(r)}$ for all $t \in \tilde{D}$.

First, assume W satisfies Normalisation$_{r,s}$. Then $W = s\tilde{W} + r$ for some normalised $\tilde{W}$. Note that $\tilde{W} = (W - r)/s$ and that $Rg(\tilde{W}) = \{(t-r)/s : t \in D\} = \tilde{D}$. We have $\tilde{\phi} \circ \tilde{W} = W^*$, since

$$\tilde{\phi} \circ \tilde{W} = \tilde{\phi} \circ [(W - r)/s)] = \phi \circ W = W^*.$$

Since $\tilde{W}$ is normalised and since $W^* = \tilde{\phi} \circ \tilde{W}$ where $\tilde{\phi}$ is smooth with $\tilde{\phi}(0) = 0$ and $\tilde{\phi}'(0) = 1$, W^* is also normalised.

Conversely, assume W^* is normalised. Since ϕ is invertible, so is $\tilde{\phi}$ $(= \phi(s \times \cdot + r))$. Further, as $\tilde{\phi}$ is the composition of ϕ with the mapping $t \mapsto st + r$, ϕ^{-1} is the composition of the latter mapping with $\tilde{\phi}^{-1}$, i.e., $\phi^{-1} = s\tilde{\phi}^{-1}(\cdot) + r$. We thus have $W = \phi^{-1}(W^*) = s\tilde{\phi}^{-1}(W^*) + r$. To show that W satisfies Normalisation$_{r,s}$, it is thus sufficient to show that $\tilde{\phi}^{-1}(W^*)$ is normalised. This follows from the fact that W^* is normalised and the fact that $\tilde{\phi}^{-1}$ is smooth with $\tilde{\phi}^{-1} > 0$, $(\tilde{\phi}^{-1})(0) = 0$ and $(\tilde{\phi}^{-1})'(0) = 1$. The second fact holds because $\tilde{\phi}$ is smooth with $\tilde{\phi}' > 0$, $\tilde{\phi}(0) = 0$ and $\tilde{\phi}'(0) = 1$.

(d) By assumption, there are $a > 0$ and $b \in \mathbb{R}$ such that $W^* = a\tau(W) + b$, or equivalently $W = \tau^{-1}((W^* - b)/a)$.

First let W satisfy CIRA$_\tau$. To show that W^* satisfies CIRA, fix a $\Delta > 0$ and prospects $p, q \in \mathcal{P}$ with equivalent welfare w.r.t. W^* denoted w_p^* resp. w_q^*, and assume $p(W^* = w) = q(W^* = w + \Delta)$ for all $w \in \mathbb{R}$. We must show that $w_q^* = w_p^* + \Delta$. Since $W^* = a\tau(W) + b$, the prospects p and q have equivalent welfare w_p resp. w_q w.r.t. W satisfying $w_p^* = a\tau(w_p) + b$ resp. $w_q^* = a\tau(w_q) + b$. For all $t \in \mathbb{R}$, we have $p(\tau(W) = t) = q(\tau(W) = t + \Delta/a)$, because

$$\begin{aligned} &p(\tau(W) = t) = p(a\tau(W) + b = at + b) = p(W^* = at + b) \\ &q(\tau(W) = t + \Delta/a) = q(a\tau(W) + b = at + b + \Delta) = q(W^* = at + b + \Delta) \end{aligned}$$

and because $p(W^* = w) = q(W^* = w + \Delta)$ for all $w \in \mathbb{R}$. We can now apply CIRA$_\tau$ to W and to Δ/a (rather than Δ). This yields $\tau(w_q) = \tau(w_p) + \Delta/a$. Thus $a\tau(w_q) + b = a\tau(w_p) + b + \Delta$, i.e., $w_q^* = w_p^* + \Delta$.

Conversely, suppose W^* satisfies CIRA. To show that W satisfies CIRA$_\tau$, note first that $Rg(W) \subseteq D$ by assumption. Next, consider a $\Delta > 0$ and prospects $p, q \in \mathcal{P}$ with equivalent welfare w_p resp. w_q, and assume $p(\tau(W) = t) = q(\tau(W) = t + \Delta)$ for all $t \in \mathbb{R}$. We prove that $\tau(w_q) = \tau(w_p) + \Delta$. As $W^* = a\tau(W) + b$, p and q have equivalent welfare w.r.t. W^* given by $w_p^* = a\tau(w_p) + b$ resp. $w_q^* = a\tau(w_q) + b$. For all $w \in \mathbb{R}$, we

have $p(W^* = w) = q(W^* = w + a\Delta)$, because

$$p(W^* = w) = p(a\tau(W) + b = w) = p\left(\tau(W) = (w - b)/a\right)$$
$$q(W^* = w + a\Delta) = q(a\tau(W) + b = w + a\Delta) = q\left(\tau(W) = (w - b)/a + \Delta\right)$$

and because $p(\tau(W) = t) = q(\tau(W) = t + \Delta)$ for all $t \in \mathbb{R}$. So, by CIRA applied to W^* and to $a\Delta$ (rather than Δ), $w_q^* = w_p^* + a\Delta$, i.e., $a\tau(w_q) + b = a\tau(w_p) + b + a\Delta$. Thus, $\tau(w_q) = \tau(w_p) + \Delta$. ■

Proof of Theorem 2. Assume $\succeq$ is regular and broad-ranging, and consider the generalised hypotheses CIRA_τ, Full-range_D and $\text{Normalisation}_{r,s}$ for given D, τ, r, s.

1. In this part, we fix a well-behaved welfare function W satisfying the generalised hypotheses, and we prove that W has the specified form. By Lemma 10, the transformed welfare function

$$W^* = (\tau \circ W - \tau(r))/(s\tau'(r)) \tag{1}$$

is well-behaved and satisfies the original hypotheses CIRA, Full-range and Normalisation. So, by Theorem 1, $W^* = \log(\rho_\succeq U + 1)/\rho_\succeq$, where U is the (unbounded) normalised VNM representation of $\succeq$, and $\rho_\succeq$ is as given in Theorem 1. Defining ρ as in Theorem 2, and noting that $\rho_\succeq = \rho s\tau'(r)$, we obtain

$$W^* = \log(\rho s\tau'(r)U + 1)/(\rho s\tau'(r)). \tag{2}$$

By (1) and (2),

$$(\tau \circ W - \tau(r))/(s\tau'(r)) = \log(\rho s\tau'(r)U + 1)/(\rho s\tau'(r)).$$

Solving this equation for W yields $W = \tau^{-1}(\log(\rho s\tau'(r)U + 1)/\rho + \tau(r))$, as claimed. Q.e.d..

2. In this part, we show that the welfare function

$$W = \tau^{-1}(\log(\rho s\tau'(r)U + 1)/\rho + \tau(r)), \tag{3}$$

with U and ρ defined as in Theorem 2, is well-behaved and satisfies CIRA_τ, Full-range_D and $\text{Normalisation}_{r,s}$. By Lemma 10, this is the case if the transformed welfare function W^* defined by (1) is well-behaved and satisfies CIRA, Full-range and Normalisation. By plugging the expression defining W into the one defining W^*, and then simplifying, one obtains $W^* = \log(\rho s\tau'(r)U + 1)/(\rho s\tau'(r)) = \log(\rho_\succeq U + 1)/\rho_\succeq$, where $\rho_\succeq$ is as in Theorem 1, or equivalently $\rho_\succeq = \rho s\tau'(r)$. So, by Theorem 1, W^* is well-behaved and satisfies the three original hypotheses, as desired. ■

References

Abdellaoui, M., Barrios, C., Wakker, P. (2007) Reconciling introspective utility with revealed preference: Experimental arguments based on prospect theory, *Journal of Econometrics* 138(1): 356–378

Adler, M. D. (2012) *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis*, Oxford University Press.

Adler, M. D. (2016). Extended Preferences, in M.D. Adler, M. Fleurbaey (eds.), *Oxford Handbook of Well-Being and Public Policy*, Oxford University Press

Adler, M. D. (2019) *Measuring Social Welfare: An Introduction*, Oxford University Press

Adler, M. D. (2025) *Risk, Death, and Well-Being: The Ethical Foundations of Fatality Risk Regulation*, Population-Level Bioethics (New York, NY, 2025; online edn, Oxford Academic)

Arrow, K. (1951) *Social Choice and Individual Values*, New York: John Wiley & Sons.

Arrow, K. (1965) The theory of risk aversion. In: *Aspects of the Theory of Risk Bearing.* Helsinki: Yrjo Jahnssonin Saatio. Reprinted in: K. J. Arrow, *Essays in the Theory of Risk Bearing.* Amsterdam: North-Holland, 1971

Baccelli, J. (2018) Risk Attitudes in Axiomatic Decision Theory—A Conceptual Perspective, *Theory and Decision* 84: 61-82

Baccelli, J. (2024) Ordinal utility differences, *Social Choice and Welfare* 62: 275–287

Baccelli, J., Mongin, P. (2016) Choice-Based Cardinal Utility, *Journal of Economic Methodology* 23(3): 268–288

Basu, K. (1982) Determinateness of the utility function: revisiting a controversy of the thirties, *Rev. Econom. Stud.* 49: 307–311

Baumol, W. J. (1958) The Cardinal Utility Which is Ordinal, *The Economic Journal* 68: 665

Bell, D.E., Raiffa H. (1988) Marginal value and intrinsic risk aversion. In: Bell D.E., Raiffa H., Tversky, A. (eds) *Decision Making: Descriptive, Normative, and Prescriptive Interactions*, Cambridge University Press, pp.384–397

Bernoulli, D. (1738) Specimen Theorize Naval de Mensura Sortis, *Commentarii Academiae Scientiarum Imperialis Petropolitanae* V: 1730–1731 [translated in Sommer, L. (1954) Exposition of a New Theory on the Measurement of Risk, *Econometrica* 22(1): 22–36]

Bossert, W., Weymark, J. (2004) Utility in social choice. In: Barberà, S., Hammond P., Seidl, C. (eds.) *Handbook of utility theory, vol. 2: Extensions*, Kluwer, Dordrecht, pp. 1099-1177

Broome, J. (1991) *Weighing Goods*, Oxford: Blackwell

Broome, J. (1991a) 'Utility', *Economics and Philosophy* 7(1): 1–12

Broome, J. (2004) *Weighing Lives*, Oxford University Press

Buchak, L. (2013) *Risk and rationality*, Oxford University Press

Bykvist, K. (2021) Taking values seriously, *Synthese* 199: 6331–6356

Chambers, C, P., Echenique, F. (2012) When does aggregation reduce risk aversion? *Games and Economic Behavior* 76(2): 582–595

Chambers, C. P. (2012) Inequality aversion and risk aversion, *Journal of Economic Theory* 147(4): 1642–1651

Chiappori, P.-A., Paiella. M. (2011) Relative risk aversion is constant: evidence from panel data, *Journal of the European Economic Association* 9(6): 1021–1052

Cibinel, P. (2025a) Welfare and autonomy under risk, *Philosophy and Phenomenological Research* 110: 526–551

Cibinel, P. (2025b) The Risk-Priority View, working Paper, Princeton University

Dietrich, F., Jabarian, B. (2022) Decision under normative uncertainty, *Economics and Philosophy* 38(3): 372–394

Dhillon, A., Mertens, J.-F. (1999) Relative utilitarianism: An improved axiomatization, *Econometrica* 67: 471–498

Ellsberg, D. (1954) Classic and Current Notions of "Measurable Utility", *The Economic Journal* 64: 528–556

Elminejada, A., Havraneka, T., Irsovaa, Z. (2022) Relative Risk Aversion: A Meta-Analysis, Working Paper, Charles University, Prague

Fishburn, P. C. (1989) Retrospective on the Utility Theory of von Neumann and Morgenstern, *Journal of Risk and Uncertainty* 2: 127–157

Fleurbaey, M. (2010) Assessing risky social situations, *Journal of Political Economy* 118: 649–680

Fleurbaey, M., Blanchet, D. (2013) *Beyond GDP: Measuring Welfare and Assessing Sustainability*, Oxford University Press

Fleubay, M., Maniquet, F. (2011) *A Theory of Fairness and Social Welfare*, Cambridge University Press

Fleurbaey, M., Mongin, P. (2005) The news of the death of welfare economics is greatly exaggerated, *Social Choice and Welfare* 25: 381–418

Fleurbaey, M., Mongin, P. (2016) The utilitarian relevance of the aggregation theorem, *American Economic Journal: Microeconomics* 8(3): 289–306

Fleurbaey, M., Zuber, S. (2021) Fair Utilitarianism, American Economic Journal: Microeconomics 13(2): 370–401

Grant. S., Kajii, A., Polak, B., Safra, Z.. (2010) Generalized utilitarianism and Harsanyi's impartial observer theorem, *Econometrica* 78(6): 1939–71

Greaves, H. (2017) A reconsideration of the Harsanyi-Sen-Weymark debate on utilitarianism, *Utilitas* 29(2): 175–213

Harsanyi, J. (1955) Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility, *J Political Economy* 63: 309–21

Harsanyi, J. (1978) Bayesian decision theory and utilitarian ethics, *American Economic Review* 68: 223-228

Hausman, D. (2012) Preference, Value, Choice, and Welfare, *Cambridge University Press*

Isbell, J.R. (1959) Absolute games, in A.W. Tucker and R.D. Luce (eds.) *Contributions to the Theory of Games*, Vol. IV, Princeton University Press

Karni, E., Weymark, J. (2024) Impartiality and Relative Utilitarianism, *Social Choice and Welfare* 63(1): 1–18

Köbberling, V. (2006) Strength of preference and cardinal utility, *Econom. Theory* 27(2): 375–391

Krantz, D. H., Luce, R. D., Suppes, P., Tversky, A. (1971) Foundations of Measurement. In: *Additive and Polynomial Representations*, vol. 1, Academic Press, New York

McCarthy, D., Mikkola, K., Thomas, T. (2020) Utilitarianism with and without expected utility, *Journal of Mathematical Economics* 87: 77–113

Mongin, P., Pivato, M. (2016) Social Evaluation under Risk and Uncertainty. In: M. D. Adler and M. Fleurbaey (eds) *The Oxford Handbook of Well-Being and Public Policy*, Oxford, U.K.

Morreau, M., Weymark, J. (2016) Measurement Scales and Welfarist Social Choice, *Journal of Mathematical Psychology* 75: 127–136

Nebel, J. M. (2021) Utils and Shmutils, *Ethics* 131: 571–599

Nebel, J. M. (2022) Aggregation Without Interpersonal Comparisons of Well-Being, *Philosophy and Phenomenological Research* 105: 18–41

Nebel, J. M. (2023) The Sum of Well-Being, *Mind* 132: 1074–1104

Nebel, J. M. (2024a) Ethics without numbers, *Philosophy and Phenomenological Research* 108: 289–319

Nebel, J. M. (2024b) Extensive measurement in social choice, *Theoretical Economics* 19: 1581–1618

Nissan-Rozen, I. (2015) Against Moral Hedging. *Economics and Philosophy* 31: 1-21

Pivato, M. (2013) Multiutility representations for incomplete difference preorders, *Mathematical Social Sciences* 66(3): 196–220

Pratt, J.W. (1964) Risk Aversion in the Small and in the Large, *Econometrica* 32: 122–136. https://doi.org/10.2307/1913738

Samuelson, P. A. (1947) *Foundations of Economic Analysis*, Harvard University Press, Cambridge, MA

Segal, U. (2000) Let's agree that all dictatorship are equally bad, *Journal of Political Economy* 108: 569–589

Sen, A. (1970) *Collective Choice and Social Welfare*, San Francisco: Holden-Day.

Sen, A. (1977) Non-linear social welfare functions: A reply to Professor Harsanyi. In R. E. Butts & J. Hintikka (Eds.), Foundational problems in the special sciences, (Vol. 2, 297–302). Springer

Sen, A. (1985) *Commodites and Capabilities*. North-Holland, Amsterdam

Shapley, L. (1975) Cardinal utility comparisons from intensity comparisons, Tech. Rep. R-1683-PR, Rand Corporation

Stigler, G., Becker, G. (1977) De Gustibus non est disputandis, *American Economic Review* 67(2): 76–90

von Neumann, J., Morgenstern, O. (1944) *Theory of Games and Economic Behavior*, Princeton University Press

Wakker, P. (1988) The algebraic versus the topological approach to additive representations. *J. Math. Psych.* 32(4): 421–435

Wakker, P. (1989) *Additive Representations of Preferences*. Kluwer, Dordrecht.

Wakker, P. (2010) *Prospect Theory: for Risk and Ambiguity*, Cambridge University Press

Weymark, J. (1991) A reconsideration of the Harsanyi–Sen debate on utilitarianism. In Elster, J., Roemer, J. E. (eds.) *Interpersonal Comparisons of Well-Being*, Cambridge University Press, pp. 255

Weymark, J. (2005) Measurement Theory and the Foundations of Utilitarianism, *Social Choice and Welfare* 25: 527–555