From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Philosophy

2025

Individual Risks and their Social Outcomes

Cibinel, Pietro

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp015138jj35s

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

Individual Risks and their Social Outcomes

Pietro Cibinel

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY

THE DEPARTMENT OF PHILOSOPHY

Adviser: Lara Buchak

Adviser: Jacob M. Nebel

September 2025

© Copyright by Pietro Cibinel, 2025. All rights reserved.

Abstract

This dissertation presents and defends an ethical outlook based on two guiding ideas: first, there is a plurality of rationally permissible attitudes towards risks that involve our welfare or well-being; and second, an ethical deliberator should defer to rational individuals' attitudes towards risk when assessing policies with uncertain consequences that affect them. John Harsanyi famously argued that deferring to people's preferences under risk leads to utilitarianism. Amartya Sen and John Weymark pointed out that Harsanyi's argument relies on a contentious interpretation of decision theory under risk—according to which, effectively, only one attitude towards risks involving our well-being, that of risk neutrality, is rationally permissible. As the Sen-Weymark critique clarifies, Harsanyi's stance is untenable if our quantitative notion of well-being is independent of preferences formed in conditions of risk.

Chapter 1 argues that this is indeed the case: our quantitative notion of well-being depends on comparisons between outcomes with no uncertainty. It defends a method for deriving quantities of well-being based on conjoint measurement between objective goods. Chapter 2 explores the space of possibilities that our two guiding ideas leave us with; it shows that under weak assumptions they are incompatible with welfarism, the view that all we need to know in order to evaluate outcomes is how people's well-being is distributed. It then presents a dilemma for non-welfarist views: they either entail that risk attitudes affect the moral significance of well-being in contexts of zero risk, or they struggle to vindicate the thought that improving everyone's well-being is an important ethical concern. Chapter 3 sides with the first horn of this dilemma. It offers a defense of a position I call the Risk-Priority View: on this theory, the ethical import of an individual ending up with a certain amount of well-being is affected by that individual's risk attitude. Roughly, more risk-averse individuals get priority in dire circumstances, while more risk-seeking individuals get priority in favorable ones.

Acknowledgments

I will first thank a number of people in connection to this dissertation in general, and to my time in graduate school, and then give chapter-specific acknowledgments.

I have so many reasons to thank my advisers, Lara Buchak and Jake Nebel, that the task now before me overwhelms me, and I am struck by an acute case of writer's block. It is a testament to their advising ability that I rarely suffered from this predicament for long while writing the dissertation itself. They have been the perfect team of advisers for a dissertation on Harsanyi's aggregation theorem, where issues of well-being measurement and distribution seamlessly combine with questions of risk-taking.

Lara encouraged me to pursue formal ethics early on. Her philosophical acumen, theory-building aspirations, and capacity for incisive criticism have accompanied me ever since, deeply influencing the trajectory of my work. I am grateful for her thoughtful support, especially at critical moments. Her work has been a constant inspiration, at times friend, at times foe, throughout graduate school. I am sorry to say I ultimately had to give up REU theory.

Jake taught me so much. He helped me enormously in giving shape to my inchoate ideas, and through this process showed me the philosophical value of attention to detail. His countless suggestions—big and small—improved the dissertation beyond what I can convey. I actually could convey that, if I shared the first draft of any of my chapters... I owe to him the way I think about (the structure of) value and what matters.

Lara and Jake's investment in my project motivated me to figure it all out for the sake of at least two other people; their good judgment and good humor made writing the dissertation so much easier, and pleasant.

I next want to thank Sarah McGrath, for being a reader of my dissertation and providing in-depth, thought-provoking comments; I also thank her for the many karaokes and guitar pulls, which I hope will continue into the future. Of the (current or past) Princeton Philosophy faculty, my gratitude also goes to: Adam Elga, for being an examiner of my dissertation and for discussing some of its ideas in their infancy—also, co-teaching PHI 371 has been a highlight of my PhD, and so valuable in learning how to teach; Harvey Lederman, for being willing to follow my work after leaving Princeton—he helped me greatly on two chapters, and has been a source of guidance throughout graduate school; Liz Harman, for her professional advice and feedback on one chapter; Tom Kelly, for being my first-year adviser and a most friendly presence in the department thereafter. In my five years at Princeton, I have benefited much from philosophical discussion with all of these faculty members, as well as with David Builes, Johann Frick, Mark Johnston, Gideon Rosen, and Peter Singer.

I would now like to thank some more academics and academic organizations. In Marcus Pivato and Franz Dietrich, I found two incredibly receptive and supportive cross-disciplinary interlocutors; I thank them for their encouragement. I was lucky enough to meet some of the protagonists of my dissertation: I thank John Broome and John Weymark for showing interest in my work, and gently steering it in the right direction. I want to thank the Global Priorities Institute, and its members, for hosting me for a month in the summer of 2023 and a week in the summer of 2024 (in connection to a Global Priorities Fellowship): it is not often that your average lunch conversation is about the meaning of vNM utility or the ex ante Pareto principle—unless you're at GPI. There, I benefited especially from talking to Gustav Alexandrie, Tomi Francis, Richard Pettigrew, Katie Steele, Teruji Thomas, and Timothy L. Williamson. I am also grateful to Calvin Baker, Chris Bottomley, Kieran Britt, Will Combs, Ryan Doody, Jeremy Khoo, Kacper Kowalczyk, Sebastian Liu, Michele Odisseas Impagnatiello, Aidan Penn, Richard Roth, Jeff Russell, Weng Kin San, and Margaret Shea for exciting philosophical conversations, and to audiences at the University of Oxford (2024 Global Priorities Workshop), Paris School of Economics (2024 Meeting of the Society for Social Choice and Welfare), Texas A&M, University of Southern California, UNC Chapel Hill, and LSE

for useful feedback. I thank the UCHV for a Graduate Prize Fellowship, and the Princeton Graduate School for a Porter Ogden Jacobus Fellowship.

Calvin Baker, Chris Bottomley, Anusha Celly, Jeremy Khoo, Sebastian Liu, João Marques Carvalho, Margaret Shea, Giulia Weißmann: thank you for sharing part or all of the journey of grad school together, and for making it lighthearted. Thanks especially to Chris for the many hours at the Yankee Doodle talking philosophy and playing shuffleboard.

Carlo Alberto, Filiberto, Ludovico and "il solito," José, Silvia: thank you for still being my friends.

A few more people or entities that I would like to thank include: Caleb Brown and the Brown-Buchak family; Keisha Craig, Anna Faiola, Nancy Groll, and Brittany Masterson; nonna Diana; Nuwan; Piero; the Shea family; Deejay Chiama Italia; the Princeton Public Library.

For everything else, I thank my family—Dante, Monica, Elisa, Silvia—and Margaret Shea: for being with me every day, for helping me take things easy, for making life good, and so on.

Now for chapter-specific acknowledgments. Lara and Jake gave me extensive feedback on all the three core chapters, looked at multiple drafts, and helped me improve the papers tremendously; Sarah and Jake acted as official readers, posing important questions about the dissertation that will inform my thinking going forward. I thank Lara, Jake, and Sarah here and omit their names below.

(**Chapter 1**) I would like to thank Chris Bottomley, John Broome, Adam Elga, Sebastian Liu, Richard Roth, Weng Kin San, and Timothy L. Williamson for helpful discussion. Thanks to the Princeton Dissertation seminar participants for exchanges and comments on a draft. I'm grateful to audiences at Princeton, Texas A&M, University of Southern California, UNC Chapel Hill, and LSE for stimulating Q&A's and follow-up discussions on a talk version of this paper.

(**Chapter 2**) This chapter is published as Cibinel (2025) in *Philosophy and Phenomenological Research*. I am grateful to Harvey Lederman, Fanhao Meng, Aidan Penn, and two referees of *Philosophy and Phenomenological Research* for written comments on drafts of the paper, and to Adam Elga and Elizabeth Harman for reading a draft and providing oral feedback. The Princeton Philosophy dissertation seminar participants read a draft and provided helpful suggestions. My thanks also go to Chris Bottomley, Franz Dietrich, Ryan Doody, Brendan Kolb, Kacper Kowalczyk, Richard Pettigrew, Marcus Pivato, Richard Roth, Weng Kin San, Margaret Shea, Katie Steele, Teruji Thomas, John Weymark, Timothy L. Williamson, and Qichen Yan for conversations or correspondence on the topic of the paper. Finally, I would like to thank participants of the UCHV Fellowship dissertation seminar for questions on a talk based on this paper.

(**Chapter 3**) I would like to thank Chris Bottomley, Richard Bradley, Franz Dietrich, Adam Elga, Kacper Kowalczyk, Marcus Pivato, Weng Kin San, Margaret Shea, Lingzhi Shi, Teruji Thomas, John Weymark, and Qichen Yan for beneficial discussion, and Harvey Lederman for that as well as for very useful written comments. I am also indebted, for constructive feedback, to a number of anonymous referees, and to audiences at the 2024 Global Priorities Workshop and 2024 Meeting of the Society for Social Choice and Welfare, where a talk version of this paper was presented.

(**Conclusion**) I would like to thank Ralph Wedgwood, Martin Peterson, Alex Worsnip, and especially Jeff Russell for illuminating conversations on the topics I address in the Conclusion.

Contents

Abstract 3

Introduction 9

1 Independent Aspects of Well-Being 14

2 Welfare and Autonomy Under Risk 49

3 The Risk-Priority View 82

Conclusion 120

Bibliography 129

Introduction

Consider three options: A is better for you than B, and B in turn better than C. These options could be anything from pasta dishes to amounts of money to ways you could spend your evening—things that bear on your welfare or well-being. You are faced with a choice between B for sure and a coin flip that gives you a 50% chance of A and 50% chance of C. What do you prefer? We can call preferences of this kind *risk-preferences*, in that they involve options whose outcome is uncertain, and therefore reveal your inclination to take risks. Suppose you are indifferent between B for sure and the coin flip on A and C. Does this mean that the amount of welfare B would give you over C equals the amount A would give you over B, as in the figure below?

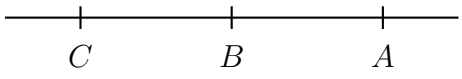


Figure I.1

The fact that you are indifferent does mean that the reason in favor of B and against the coin flip—that the latter gives you C instead of B with 50% probability—is exactly balanced out by the reason against B and in favor of the coin flip—that the latter gives you A instead of B with 50% probability. So it might seem natural to think: since the improvement from C to B counts the same as that from B to A in determining your risk-preference, the two are welfare improvements of the same *size*.[1]

Suppose instead you prefer B. Through a similar argument, we might reason our way to the conclusion that the amount of welfare B would give you over C now *exceeds* the amount A would give you over B, as illustrated in the next figure.

[1] I am here paraphrasing and adapting the argument in Broome (1991b, pp. 146-147); on Broome's version of the argument, what matters is whether B for sure is better for you than the coin flip.

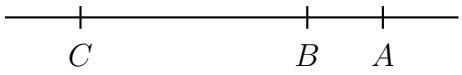


Figure I.2

Suppose finally you prefer the coin flip. Again, we might be tempted to reach the analogous conclusion that the amount of welfare B would give you over C is here less than the amount A would give you over B, as in this final figure.

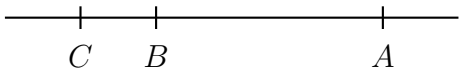


Figure I.3

This line of thought expresses a position that we can call *risk neutrality* in welfare: your risk-preferences are such as to maximize expected welfare. That is, welfare is understood as a quantity that you maximize the expectation of, where the expectation of a quantity is given by a weighted sum of the values the quantity takes in each possible outcome, each value weighted by the probability of the given outcome.

Risk neutrality plays a central role in John Harsanyi's powerful argument for utilitarianism, the view that a social outcome should be evaluated by a sum or average of each individual's welfare.[2] Harsanyi came to accept a form of risk neutrality in welfare by equating welfare with preference satisfaction and by interpreting risk-preferences as eliciting intensity of preference.[3] On this view, for example, if you prefer B for sure over a coin flip between A and C, you thereby more intensely desire B over C than A over B. Given this understanding of the relationship between welfare and risk-preferences, Harsanyi's (1955) renowned social aggregation theorem shows—with some finessing (Harsanyi 1977, 1982; Weymark 1991)—that utilitarianism is entailed by the standard axioms of decision theory, applied both at the individual and social level, plus the requirement that an ethical deliberator should defer to people's risk-preferences.

Amartya Sen (1976), (1977) and John Weymark (1991) challenged Harsanyi's argument by ques-

[2]Harsanyi (1977), (1982) in fact had two arguments for utilitarianism, one based on his "equiprobability model" (1953; 1982, p. 44) and the other on his social aggregation theorem (1955; 1982, p. 48). My interest will be in the latter argument, though the two are related (for example, certain criticisms apply to both). It is also worth pointing out that Harsanyi preferred a rule utilitarian position over act utilitarianism (1977, pp. 63-64).

[3]In interpreting Harsanyi as sometimes using "utility" to refer to a person's good—or, as I shall more often call it, their welfare or well-being—I follow Broome (1991b, pp. 58-59, n. 19).

tioning his interpretation of the significance of risk-preferences for quantitative facts concerning welfare. Consider again our example, and suppose we grant the idea that welfare should be equated with preference satisfaction. The following scenario seems possible: although you have an equally strong desire for B over C as for A over B, you want to be careful, and when faced with a choice between either B for sure or a coin flip on A and C, you would rather rest content with B for sure. Abstracting now from the preference satisfaction theory of welfare, the following scenario also seems possible: although the welfare difference between B and C is the same as that between A and B, you want to be careful, and when faced with a choice between either B for sure or a coin flip on A and C, you would rather rest content with B for sure.

If these scenarios are possible, your preferences in them seem perfectly reasonable: there seems to be nothing irrational in the kind of risk aversion they exemplify.[4] Harsanyi's position might have been that these scenarios are not in fact possible. On this version of the risk neutrality assumption, there is a constitutive connection between risk-preferences and preference intensities, and thus between risk-preferences and amounts of welfare. The idea is that preferring B for sure over the coin flip involving A and C is *what it means* to say that you prefer B to C more strongly than A to B, and thus that there is a higher welfare difference between B and C than between A and B (Broome 1991b applies this point directly to amounts of welfare; Broome 2004 and Greaves 2017 claim that our quantitative notion of welfare is vague and that this strategy provides a way to make it precise). The question then becomes: is our notion of welfare really so impoverished as to not make any quantitative sense outside the context of risk-preferences?

If we could somehow derive amounts of welfare independently of risk-preferences, risk neutrality in welfare would appear dubious: preferring B for sure to the coin flip, even though the amount of welfare in B is less than halfway between the amounts in C and A, would be a possible and reasonable stance for you to take. In this dissertation, I corroborate and build upon the the Sen-Weymark critique of Harsanyi, by rejecting risk neutrality via a defense of an independent quantitative mea-

[4] As we will see later, this kind of risk aversion is perfectly compatible with orthodox decision theory, and need not be of the kind espoused by Buchak (2013).

sure of welfare. I then, however, retain Harsanyi's idea that an ethical deliberator should defer to people's risk-preferences. And I proceed to investigate the consequences of these two premises for ethics.

Chapter 1 is devoted to arguing that we can indeed make sense of amounts of welfare independently of risk-preferences. Quantities of welfare are based on tradeoffs between objective goods. Given this conception of quantitative welfare comparisons, I believe we should accept permissivism about risk-preferences for welfare: reasonable people can differ in their evaluation of risks in welfare. Suppose that the welfare improvement from C to B is somewhat smaller than that from B to A. Should you prefer the coin flip to B for sure? It seems your preference could rationally go either way. **Chapter 2** (published as Cibinel 2025) explores the consequences of accepting this idea for ethics. It argues that, if we want to vindicate the thought that an ethical deliberator should defer to people's attitudes to risk, we must abandon welfarism, the popular view according to which all we need to know in order to evaluate an outcome is how people's welfare is distributed.[5] The chapter, however, also presents a dilemma for non-welfarist views: they either have to accept that risk attitudes affect the moral significance of welfare in contexts where there is no risk, or else they struggle to accommodate the intuition that welfare is an important value to be promoted. **Chapter 3** defends the first horn of this dilemma. It lays out in greater detail, and then defends, a view introduced in **Chapter 2**. On this proposal, which I call the Risk-Priority View, the social or ethical value of a person ending up at a given welfare level depends on that person's attitude to risk. This is the view that a suitable reinterpretation of Harsanyi's social aggregation theorem lands us on, once the Sen-Weymark critique has been taken into account.

The view I end up proposing is no doubt counterintuitive. The Sen-Weymark critique of Harsanyi

[5]Note that, here and in **Chapter 2**, I define welfarism as a thesis concerning outcomes only. Some authors (e.g., Blackorby, Donaldson, and Weymark 2008, discussed in **Chapter 2** and **Chapter 3**) develop models in which welfare also attaches to options with uncertain outcomes, and understand welfarism as the thesis that all we need to know in order to evaluate alternatives (whether outcomes or options with uncertainty) is how people's welfare is distributed. Their version of the welfarist tenet entails the form of welfarism I focus on. My choice relies on (i) doubting that it makes sense to think of welfare as attaching to options with uncertainty, and (ii) thinking that if it does make sense, we should leave open that different measurability/comparability assumptions apply for welfare on outcomes *vs.* welfare on options with uncertainty, so that it may anyway be best to keep the two notions apart.

is at least partly informed by a tradition that rejects *any* use of risk-preferences for the evaluation of social outcomes. Kenneth Arrow (1951) famously remarked that to think risk-preferences are relevant when

> [...] we are interested primarily in making a social choice among alternative policies in which no random elements enter [...] would be to assert that the distribution of the social income is to be governed by the tastes of individuals for gambling. (p. 10)

It is instructive to consider Harsanyi's reply to this kind of worry. He says:

> [...] our utilitarian theory will tend to assign higher social priorities to those individual desires for which people are willing to take considerable risks in order to satisfy them. But this is surely as it should be. Other things being equal, we *should* give higher social priorities to intensely felt human desires; and one indication that somebody feels strongly about a particular desired objective is his willingness to take sizable risks to attain it. (Harsanyi 1982, p. 53, emphasis in original)

As I see it, Harsanyi's reply is distinctively utilitarian. He is justifying the relevance of risk-preferences as follows: intensities of desire—which he takes amounts of welfare to consist in—clearly matter, and risk-preferences allow us to measure such intensities. I believe this passage suggests that, if he had somehow become convinced that intensities of desire are not reliably tracked by risk-preferences, he would have reconsidered the relevance of risk-preferences for the ethics of evaluating social outcomes.

My defense of the significance of risk-preferences is not welfarist in this way. But how else could they possibly matter? An answer awaits the reader in **Chapter 3**. To summarize my dissertation in one sentence: I side with the Sen-Weymark critique of Harsanyi in questioning the significance of risk-preferences for measuring welfare, but resist concluding that risk-preferences are therefore insignificant for social evaluation; instead, they measure the moral value of welfare.

Chapter 1

Independent Aspects of Well-Being

Abstract

Can we make sense of the idea that there are *quantities* of individual well-being? I argue for an affirmative answer, on which quantitative well-being comparisons are based on tradeoffs between prudential, final, objective goods—for example, pleasure, knowledge, and friendship.

1.1 Introduction

Our well-being or welfare is what should matter most to each of us, insofar as we are concerned with our own interest. It is what's good for us (Parfit 1984, p. 493). Suppose a friend comes to you and asks for advice on an important matter. In trying to be helpful, you think about what would be in your friend's interest: what would be best for your friend (Darwall 2002). That is, you think about their well-being. The question of this paper is whether well-being is the type of thing that can at least in principle be measured quantitatively. Are there *quantities* of individual well-being? I will argue for an affirmative answer, on which

> (**Pluralistic Tradeoffs Account**) quantitative comparisons of well-being are based on tradeoffs between objective goods.

Objective goods will be understood as the kinds of items that feature on objective list theories of well-being (Parfit 1984, p. 499).

To get a grip on our topic, notice that we can make different kinds of comparisons involving well-being.[1] First, we can rank people's well-being: we can say whether a certain person is better off than another person, and whether a certain person would be better off in one situation or in another situation.[2] Suppose, for instance, that Alex's life is filled with pleasure, meaningful relationships, and a sense of accomplishment. Brooke's life contains some sadness and dissatisfaction, but also real friendships. Charlie's life is painful throughout, containing little of what makes life worth living. Intuitively, Alex is better off than Brooke, and Brooke, in turn, better off than Charlie. These are ordinal comparisons of well-being, in that they amount to ordering people's lives, in terms of which are better from the perspective of well-being.

But perhaps we can say something more informative. Perhaps we can say that, although Brooke is better off than Charlie, she is not better off *by much*, compared to the way in which Alex is much better off than Brooke. This judgment can be represented graphically as follows.

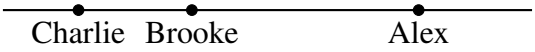


Figure 1.1

In other words, we might be inclined to say that the difference or improvement in well-being from Charlie to Brooke is smaller than the difference or improvement in well-being from Brooke to Alex. Maybe, we could say something more precise still; for instance, that the improvement in well-being from Charlie to Brooke is *one third* in size the improvement in well-being from Brooke to Alex. I will call these quantitative comparisons of well-being.

In general form, a quantitative comparison of well-being looks like this:

This given difference or improvement in well-being is greater than (or x-times as great

[1] See Adler (2019, pp. 11-15, 41-47, 124-130) for an excellent introduction to various kinds of well-being comparisons, including the ones I am about to present.

[2] This paper does not deal with *interpersonal* well-being comparisons; for illustration, I simply assume, as many philosophers do, that they can be made. If they cannot be made, then my question is still whether well-being comes in quantities, only these will now be of different *dimension* (as, e.g., mass and length) for different people (see Nebel 2022a).

as) *that* other difference or improvement in well-being.

Here, x could be a positive natural number, such as 3, but also a positive rational number, such as $\frac{3}{2}$, or potentially an irrational number, such as $\sqrt{2}$. These comparisons will be the target of this paper. I will give a theory of what explains comparisons of this kind, of what makes such comparisons possible or meaningful.[3]

A note about the connection between ordinal and quantitative comparisons: some quantitative comparisons can be made on the basis of ordinal comparisons alone, and everyone should accept that (Baccelli 2024). In the example above, where Alex is better off than Brooke and Brooke better off than Charlie, this ordinal information alone suffices to conclude that the well-being improvement from Charlie to Brooke is smaller than that from Charlie to Alex. But this ordinal information alone does not allow us to conclude anything about how the well-being improvement from Charlie to Brooke compares to that from Brooke to Alex. My concern in this paper is specifically with quantitative comparisons of the latter kind, which cannot be inferred from ordinal information alone.

This is our topic—why care about it? As it happens, there are many positions in moral philosophy that we might want to entertain, or even adopt, which presuppose quantities of well-being. I will here mention one example. On the *leximin* view in distributive ethics, if the worst off individual in one society is better off than the worst off individual in another society, then the former society is better. While this view aims to back up an admirable ideal of equality, many find it implausible upon close inspection (e.g., Parfit 2011, p. 353). If we can either benefit the worst off individual in a trifling way (which would leave them worst off), or else the second worst off individual in a substantial way, it enjoins us to choose the former option. But if the situation of the

[3]The relevant notion of meaningfulness here is the one familiar from measurement theory (see, e.g., Krantz et al. 1971, Sect. 1.2.3). In that jargon, my question is whether well-being is at least amenable to interval scale measurement. I do not deal with the more demanding requirement of ratio scale measurability (essentially, of whether there is a natural zero of well-being). See Nebel (2023) for an examination of this latter question. If well-being can be measured on a ratio scale, then it is *a fortiori* amenable to interval scale measurement; I shall note more points of contact with Nebel's paper below.

second worst off individual is not *much* different from that of the worst off individual, it seems we should choose the latter option. After all, it would allow us to benefit substantially someone who is almost worst off. Such a plausible rationale for rejecting leximin appeals to quantitative well-being comparisons: it presupposes that we can make sense of a benefit being trifling, or substantial, in comparison to another, and that two people's well-being can be very close.[4]

Moral philosophers, indeed, routinely assume the possibility of quantitative well-being comparisons.[5] But it's not clear that they are justified. Typically, statements about quantities of a certain kind are closely associated with a certain measurement procedure. For example, to determine which of two sticks is longer, we can grab a meter and see how many centimeters fit into the first stick, as well as how many fit into the second: if more fit into the first than into the second, the first is longer. But it's doubtful whether well-being can be measured in anything like this way. And indeed, some philosophers take a different stance: they think that we need to earn our right to talk about quantities of well-being (Broome 1991b). This paper shares this methodological starting point. If there are quantities of well-being, they shouldn't be taken as primitive, and need to be shown to be based on facts that we have a better grip on. I will argue that we can, in the end, vindicate moral philosophers' reliance on quantitative talk concerning well-being.

Let me clarify how I see the contribution of this paper. The most popular approach in formal ethics is the "vNM method," on which quantitative well-being comparisons are based on tradeoffs arising in conditions of risk, through the use of the von Neumann-Morgenstern (1944) representation theorem, or variants thereof (Broome 1991b, 2004; Jensen 1995; Risse 2002; McCarthy 2006, 2007; Greaves 2017; Adler 2019; while I call it the vNM method, I do not mean to attribute it to von Neumann and Morgenstern themselves). This theorem shows that under certain axioms, there exists an additive (more precisely, expectational) real-valued representation of a preference

[4]Following up on footnote 2, those that deny interpersonal comparisons may wish to reinterpret this case as providing a rationale, based on quantitative well-being comparisons, for rejecting leximin as a decision theory under ignorance.

[5]For example, both utilitarians and prioritarians require these kinds of comparisons, as do various proponents of "limited aggregation," such as Voorhoeve (2014).

relation, unique up to positive affine transformation. The interpretative option of treating this additive representation as measuring preference intensities has been around ever since the inception of the theory.[6] But arguably, it is only with Broome (1991b) that we get a full-fledged philosophical defense of the idea that decision theory under risk provides a foundation for quantities of *well-being*—with the preference relation reinterpreted as a personal betterness relation, and the focus shifting to normative, and away from psychological, theorizing.[7]

Conjoint measurement between bundles of "goods," commodities, or dimensions of some phenomenon of interest is another route, well-known to economists and measurement theorists, via which an additive real-valued representation of a preference relation, unique up to positive affine transformation, may be achieved.[8] The interpretative option of treating this additive representation as measuring preference intensities has been around ever since the inception of the theory.[9] Nonetheless, there seems to be no established reference in the *philosophical* literature for a detailed defense of this method as providing the correct foundation for quantities of well-being.[10] No doubt only in aspiration, the point of this paper is to parallel what Broome did for the vNM method, but for conjoint measurement between kinds of goods.[11]

I proceed as follows. In Section 1.2, I will present the basic idea of my account informally, with an illustration. Section 1.3 develops the account more precisely and in full generality. Sections 1.4

[6] See, e.g., Harsanyi (1977); see also Fishburn (1989) for references and a criticism of this interpretation.

[7] Broome (1991b) also explores the strategies of deriving quantities of well-being from interpersonal tradeoffs and tradeoffs across times. I plan to address the relationship between my account and these three alternative approaches in a companion paper (see the Conclusion of this dissertation for some initial remarks).

[8] See Suppes and Winet (1955), Debreu (1960), Luce and Tukey (1964), Krantz et al. (1971), Sarin and Dyer (1979), Bell and Raiffa (1988), Wakker (1989), Keeney and Raiffa (1993). As I clarify in the Appendix, what will matter for my project is the fact that conjoint measurement shows the existence of an additive representation; this should not be confused with the idea, to be rejected, that the additive representation is somehow privileged (see Krantz et al. 1971; Weymark 2005; Nebel 2024).

[9] See, e.g., Suppes and Winet (1955, p. 260) and Bell and Raiffa (1988, pp. 388-395). See footnote 24 below for more on the relationship between this claim and the account I will develop in this paper.

[10] For some relevant work, however, see Griffin (1986, ch. 6), Oddie (2001), Adler (2019, pp. 71, 268), and Stefánsson (2023)—these references are discussed in footnotes below.

[11] Another connection worth mentioning at the start is with the "equivalence approach," a method for making interpersonal comparisons of well-being that likewise appeals to tradeoffs between different "goods," such as money and health. See Adler and Decanq (2022) for a helpful summary; as they emphasize (p. 148), the equivalence approach is compatible with quantitative well-being comparisons but on its own neither supports nor threatens their meaningfulness.

and 1.5 provide an extended argument for the account. Section 1.6 tackles a residual issue. Section 1.7 concludes the paper.

1.2 Tradeoffs between Objective Goods

On the view I will develop—the **Pluralistic Tradeoffs Account**—quantities of well-being arise from tradeoffs between different goods of a certain kind. I have in mind the kind of goods that feature on objective list theories of well-being. These are pluralistic theories of well-being, on which well-being is constituted by a number of components or dimensions (see Parfit 1984, Fletcher 2015). To give an example, Fletcher (2013, p. 214) suggests that "[...] well-being is enhanced by: Achievement, Friendship, Happiness, Pleasure, Self-Respect, Virtue."

More important than what specific items go on the list is what properties all such items are supposed to share. Any good on the list is *prudential*, or considered insofar as it is related to the given individual's own interest, and not to other values, such as aesthetic or moral value. Any good on the list is *final*, or worth pursuing for its own sake for one's well-being. It is not just instrumentally valuable, or worth pursuing because it leads to some further thing that increases one's well-being. Finally, any good on the list is *objective*. If friendship is on the list, having genuine friends in your life thereby makes your life better, independently of your attitudes towards friendship (for example, independently of whether you believe that having friends improves your well-being).[12]

I refer to goods with these properties as *objective goods* for short. The basic idea of the **Pluralistic Tradeoffs Account** is that tradeoffs between objective goods give us a way to make sense of quantitative comparisons of well-being.[13] I now illustrate how the account works with a simple

[12]Cf. Fletcher (2016, p. 50) on "Attitude-Independence." This point may require some finessing to allow for desire or preference satisfaction to count as items on the objective list (see Adler 2019, p. 71, fn. 22; Adler and Decanq 2022, p. 163).

[13]Griffin (1986, ch. 6)—who accepts a form of pluralism—suggests that we can derive some quantitative well-being comparisons ("pockets of cardinality," p. 98) by considering situations in which leisure time is valued linearly and then use willingness to sacrifice units of leisure time as a measure of how much other things are valued. This is essentially a conjoint measurement scheme, but it does not reap the full benefits of the approach; the axiomatization I shall follow does not require any assumption of linear marginal value. Adler (2019, pp. 71-72) mentions the possibility of

case, relying here and throughout on pleasure, friendship, and knowledge as examples of objective goods. For concreteness, I will in fact assume that these are the only objective goods.

Suppose you need to make a tradeoff between pleasure and friendship. What I mean by this is that you can either get a lower level of pleasure $p1$ and higher level of friendship $f2$, or else a higher level of pleasure $p2$ and lower level of friendship $f1$. In both cases, the knowledge level is fixed at $k1$. The case is represented graphically below ($k2$, also in the figure, will be considered momentarily).[14] We will think of a life as a list, specifying how one fares with respect to each good: so, for example, a possible life is represented as $(p1, f1, k1)$.

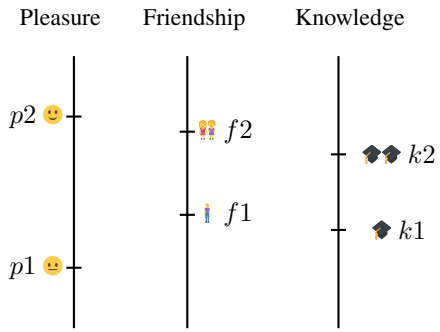


Figure 1.2

Suppose it is better for you to be at the higher level of pleasure and lower level of friendship than *vice versa*. This means that the improvement in pleasure from $p1$ to $p2$ is, in this case, worth more than the improvement in friendship from $f1$ to $f2$. Suppose next that this comparison remains true if we change the background level of knowledge from $k1$ to any other level—for example, $k2$. That

combining an objective list theory with an approach known as difference measurement. My account adds that objective goods are not only compatible with, but offer *an interpretation* of, difference measurement for well-being in terms of conjoint measurement. Stefánsson (2023) settles on a difference measurement axiomatization, but discusses an example adapted from Suppes and Winet (1955) that I interpret as illustrating the basic idea of conjoint measurement. Again, while the two axiomatizations are related, I think the axioms of conjoint measurement are justificatorily more fundamental in our context (see also footnote 23 below). Stefánsson understands the choice between different accounts of quantities of well-being as a "modelling choice" (p. 305), whereas I think it is a substantive philosophical issue, and set out to argue for my preferred account. Adler and Decanq (2022, pp. 162-163) briefly point to the possibility of combining the equivalence approach (mentioned in footnote 11) with an objective list theory, but do not remark on the advantage of this option that I will here highlight: that focusing on exhaustive bundles of objective goods instead of general bundles of commodities (e.g., money and health) allows for a normatively compelling application of conjoint measurement and consequent derivation of quantities of well-being.

[14] This illustrative device is adapted from Krantz et al. (1971, p. 254).

is, no matter what fixed level of knowledge we pick, the pleasure improvement is worth more than the friendship improvement. Notice that, so far, we are making ordinal comparisons alone. We are saying that for any level of knowledge k, a life at $(p1, f1, k)$ is worse than one at $(p1, f2, k)$, which is in turn worse than one at $(p2, f1, k)$. This will entail a certain kind of "trivial" quantitative comparison: that the well-being improvement from $(p1, f1, k)$ to $(p1, f2, k)$ is smaller than that from $(p1, f1, k)$ to $(p2, f1, k)$.

But we may also want to compare the improvement from $(p1, f1, k1)$ to $(p2, f1, k1)$ with the improvement from $(p1, f1, k2)$ to $(p1, f2, k2)$. We now ask: which of these two improvements in well-being is larger? This is a quantitative well-being comparison that cannot be derived from the ordinal ranking.[15] On the **Pluralistic Tradeoffs Account**, the fact that the $p1$-$p2$ improvement in pleasure is worth more than the $f1$-$f2$ improvement in friendship in all cases in which we need to make a tradeoff justifies the conclusion that the improvement from $(p1, f1, k1)$ to $(p2, f1, k1)$ brings about a greater well-being difference than the improvement from $(p1, f1, k2)$ to $(p1, f2, k2)$. The idea is intuitive enough. We ask: why is the former improvement in well-being greater than the latter? The following answer seems perfectly illuminating: because in cases in which we need to make a tradeoff between the two, the former always wins out. In the next section, I develop this thought into a comprehensive theory of quantitative well-being comparisons.

1.3 General Account

I will assume there is a finite number n of at least three objective goods, indexed by $N = \{1, 2, ..., n\}$. For each good i, there is a set L_i of at least two ways an individual's life could go with respect to good i: each of these is a maximal specification of how a life could go respect to the given objective good. For example, if 1 indexes pleasure, any member of L_1 will settle all questions about how a life could go with respect to pleasure. For each objective good i, there is an ordering $\succeq_i$ on L_i, encoding information about which members of L_i promote the given good

[15] More precisely, it cannot be so derived unless: $(p1, f1, k1)$ is worse than $(p1, f1, k2)$, which is worse than $(p1, f2, k2)$, which is worse than $(p2, f1, k1)$.

more.[16] For example, there will be a ranking of ways one's life can go with respect to pleasure, which tells us which of these ways contain more pleasure. We make the assumption that each of these rankings is non-trivial: for each good, there are worse and better ways for a life to go with respect to it.

As previewed, we can think of a life as a list of n entries, where each entry takes a particular member of $L_1, ..., L_n$, respectively, as its value: a specification of how the individual fares with respect to the given objective good. There is an ordering $\succeq$ on the set of all metaphysically possible lives, L, encoding ordinal well-being comparisons. We assume that a "Pareto" improvement, more of a given good at no cost to other goods, is always better. If we keep all but the level of pleasure fixed, and improve the level of pleasure, we get a better life. If that weren't the case, pleasure wouldn't be a *good*. And we assume that two lives that are equally good for all objective goods have equal well-being.

Pareto: If a life does at least as well as another with respect to all goods, it is at least as good for well-being. If, in addition, it does better with respect to some goods, it is better for well-being.[17]

The **Pluralistic Tradeoffs Account** derives quantitative comparisons from ordinal comparisons involving tradeoffs between objective goods. To make tradeoffs between goods, we need to improve some goods, and decrease others. If we want to know how some improvement in pleasure compares to some loss in friendship, we need to do the following: first, start at some reference level of pleasure and friendship; then, improve the level of pleasure and decrease that of friendship. To do this, we need a kind of independence between the goods: it should be possible to choose one way for a life to go with respect to pleasure and combine it with all sorts of ways for a life to go with respect to friendship, as well as *vice versa*. We might therefore be inclined to identify the set of all lives L with the Cartesian product $L_1 \times ... \times L_n$.

[16] An ordering is a transitive and complete relation; it follows that it is also reflexive.

[17] Hedden and Muñoz (2024) offer a compelling defense of "Pareto" principles of this kind; my own argument in favor of a more controversial principle that I shall introduce momentarily, "Normative Independence," can be understood as pushing their approach one step further.

A problem, however, now arises. Some combinations of instantiations of the objective goods will not be compatible. A way for my life to go with respect to pleasure is that I have a lot of fun with friends ($p2$). If that's how my life goes with respect to pleasure, it cannot go too badly with respect to friendship: the specification of the objective good of friendship cannot be that I have no friends ($f1$). To get around this issue, we will work at a higher level of abstraction—though in a way that is ultimately compatible with the idea that facts about well-being bottom out in which concrete lives are better for an individual.[18] Since we have an ordering for each L_i, we can extract the set $\mathcal{L}_i$ of equivalence classes induced by the symmetric part of $\succeq_i$. These are sets of all equally good ways for a life to go with respect to good i. We will call each equivalence class thus obtained a *degree* of good i (compare Nebel 2024, p. 293). For example, an exhaustive set of equally good ways for a life to go with respect to pleasure is a degree of pleasure. The idea is that differences between members of this set are evaluatively irrelevant, as far as pleasure goes. In the first instance, we will evaluate a set $\mathcal{L} \subseteq \mathcal{L}_1 \times ... \times \mathcal{L}_n$ the members of which, $\boldsymbol{a} \in \mathcal{L}$, I will call *abstract* lives, to distinguish them from the concrete lives $a \in L$ that a person might live. $\mathcal{L}$ contains all and only combinations of degrees of objective goods that are instantiated by some possible concrete life.

Then we can make good sense of a well-being ordering $\succeq$ on $\mathcal{L}$ derived from, and reflecting, ordinal well-being information on concrete lives as given by $\succeq$. Note that a total order on $\mathcal{L}_i$, $\succ_i$, can be derived from $\succeq_i$.[19] By Pareto, any two lives instantiating the same degrees of objective goods are equally good. With $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{L}$, $\boldsymbol{a} \succeq \boldsymbol{b}$ if and only if some concrete life instantiating $\boldsymbol{a}$'s degrees of the objective goods is at least as good as some concrete life instantiating $\boldsymbol{b}$'s degrees, as per $\succeq$. Notice also that a principle corresponding to Pareto for $\succeq$ follows. We provisionally assume:

Metaphysical Independence: For any possible combination of degrees of objective goods, there

[18] Dreier (1996) argues for the legitimacy of "impractical preferences," and for the claim that they have an important role to play in decision theory, understood as a theory of practical rationality. In a similar fashion, we might attempt to defend the claim that there is nothing mysterious about well-being comparisons holding between impossible lives; I do not pursue that strategy here.

[19] A total order is an antisymmetric ordering.

is a concrete life in L that instantiates that combination (that is, $\mathcal{L} = \mathcal{L}_1 \times ... \times \mathcal{L}_n$).

We will return to this assumption below, in Section 1.6. For now, notice how working with abstract lives helps us get around the initial problem.[20] A way for my life to go with respect to pleasure is that I have a lot of fun with friends ($p2$). It cannot then go too badly with respect to friendship ($f1$). This seems to prevent us from making certain tradeoffs between pleasure and friendship. What if we want to compare an improvement in pleasure from some $p1$ to $p2$ with a loss in friendship from some $f2$ to $f1$? We can work around this issue by appealing to degrees of pleasure and friendship. The comparison is then between an improvement in the degree of pleasure and one in the degree of friendship. And we can perfectly well combine a high degree of pleasure (as instantiated by $p2$) with a low degree of friendship (as instantiated by $f1$): plausibly, there will be a concrete life equally good for pleasure as the one in which you have fun with friends, where the pleasure is derived from an experience machine (Nozick 1974, p. 42), and relative to which we *can* freely change how well you fare in connection to friendship.[21]

With this problem out of the way, we can proceed to derive quantities of well-being out of $\succeq$. To this end, we impose extra structure on the pair $\langle \mathcal{L}, \succeq \rangle$. This extra structure will be familiar to those acquainted with *additive conjoint measurement* (Krantz et al. 1971; Wakker 1989; Keeney and Raiffa 1993). I will list each axiom informally and provide an example to illustrate what it says; an Appendix at the end states these axioms and the account that follows formally.

I. Normative Independence: For every subset of goods, how tradeoffs are made within that subset does not depend on how the remaining goods are arranged.

$-$*Example*: how you make tradeoffs between friendship and pleasure does not depend on what

[20] This solution adapts analogous moves by Broome (2004, p. 96), Nebel (2023, p. 1084, 1086, 1092), and San (ms). San investigates a principle, "Replacement," which is the analogue of Metaphysical Independence in the context of interpersonal and intertemporal tradeoffs.

[21] Suppose the initial concrete life is $(p1, f2, k)$, and we want to compare an improvement in pleasure from some $p1$ to $p2$ with a loss in friendship from some $f2$ to $f1$. While $(p2, f1, k)$ may not be possible, there will plausibly be some concrete life that instantiates the abstract life $(\boldsymbol{p2, f1, k})$; say, we have $p2^* \sim_1 p2$ such that the concrete life $(p2^*, f1, k)$ *is* possible. The judgment at the level of abstract lives $(\boldsymbol{p2, f1, k}) \succ (\boldsymbol{p1, f2, k})$, that the pleasure improvement is worth it, translates to $(p2^*, f1, k) \succ (p1, f2, k)$ at the level of concrete lives.

level of knowledge you are at.

II. Restricted Solvability: Arbitrarily select an abstract life $\boldsymbol{a} \in \mathcal{L}$, and also arbitrarily specify degrees of goods for all but one good i. Suppose there exists a degree of good i, $\underline{\boldsymbol{b}}_{\boldsymbol{i}}$ (respectively $\overline{\boldsymbol{b}}_{\boldsymbol{i}}$) that makes the abstract life resulting from the arbitrarily selected degrees of goods plus $\underline{\boldsymbol{b}}_{\boldsymbol{i}}$ (resp. $\overline{\boldsymbol{b}}_{\boldsymbol{i}}$) at least as bad as (resp. at least as good as) $\boldsymbol{a}$. Then there exists a degree of good i, $\boldsymbol{b}_{\boldsymbol{i}}$, that makes the abstract life resulting from the arbitrarily selected degrees of goods plus $\boldsymbol{b}_{\boldsymbol{i}}$ *exactly* as good as $\boldsymbol{a}$.

$-$*Example*: Choose an abstract life $\boldsymbol{a} \in \mathcal{L}$ and consider some particular losses in pleasure and friendship with respect to $\boldsymbol{a}$. Suppose we can find both a degree of knowledge that, coupled with those losses in pleasure and friendship, makes the resulting life at least as bad as $\boldsymbol{a}$ *and* a degree of knowledge that, coupled with those losses in pleasure and friendship, makes the resulting life at least as good as $\boldsymbol{a}$. Then we will also be able to find an improvement in knowledge that exactly compensates those two losses in pleasure and friendship: an improvement in knowledge that, together with those two losses, makes for a life exactly as good as $\boldsymbol{a}$.

III. Archimedeanism: No good is infinitely more valuable than any other.

$-$*Example*: Pleasure is not infinitely more valuable than knowledge. For a small enough loss in pleasure, we can find an improvement in knowledge that makes up for it.

If these axioms hold, two important conclusions follow (Krantz et al. 1971, p. 302, THEOREM 13): (i) for any good, each "unidimensional" improvement in that good can be assigned a number indicating how large it is, with higher numbers corresponding to larger improvements; (ii) for any good (resp. any two goods), we can define the ratio m/n between any two unidimensional improvements in that good (resp. those two goods), roughly in the sense that we can find a small enough unidimensional improvement of any good that "fits" m times into the former, and n times into the latter.[22] The *length* of an improvement is a metaphor for how much it matters in deter-

[22]This is a rough statement because such an exact fit may not exist; in that case, the result is that as we select a smaller and smaller improvement for purposes of "fitting," the ratio of the number of times it fits approximately into

mining the ordinal well-being ranking $\succeq$: we can measure its worth or the difference it makes when comparing tradeoffs just like we can measure how long a stick is. I shall also refer to an improvement's *weight* in this way. This result does not on its own amount to a general theory of quantitative well-being comparisons, but we can use it to develop one.

Suppose we select four concrete lives, a, b, c, d, with b better than a, and d better than c. We want to know:

(i) whether the difference in well-being between a and b is greater than that between c and d;

(ii) whether the difference in well-being between a and b is x-times as great as the difference in well-being between c and d.

To answer (i), we ask: does the ab collection of improvements, taken as a whole, weigh more than the cd collection of improvements? To answer (ii), we ask: does the ab collection, taken as a whole, weigh exactly x-times as much as the cd collection? On the **Pluralistic Tradeoffs Account**, if the answer to the first question is "yes," then the difference in well-being between a and b is greater than that between c and d; and if the answer to the second question is "yes," then the difference in well-being between a and b is x-times as great as that between c and d.

How do we determine whether the ab collection, taken as a whole, weighs more than the cd collection? And how do we determine whether the ab collection, taken as a whole, weighs x-times as much as the cd collection? First, we consider whether the ab and cd collections involve only unidimensional improvements proper or also unidimensional losses (that are more than compensated for, since $b \succ a, d \succ c$). In the latter case, we find lives $a^* \sim a$, $b^* \sim b$, $c^* \sim c$, and $d^* \sim d$ such that the a^*b^* and c^*d^* collections involve only unidimensional improvements proper (Restricted Solvability ensures these lives exist).

We look at the corresponding abstract lives $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}$ that a, b, c, d (or a^*, b^*, c^*, d^*) instantiate. We

the former to the number of times it fits approximately into the latter converges to m/n. This is clarified for the general account in the main text below.

try to find natural numbers m, n such that we can get from $\boldsymbol{a}$ to $\boldsymbol{b}$ via a series of m equally weighty unidimensional improvements, and from $\boldsymbol{c}$ to $\boldsymbol{d}$ via a series of n equally weighty unidimensional improvements, each equally weighty as the m improvements we use to go from $\boldsymbol{a}$ to $\boldsymbol{b}$. To get from $\boldsymbol{a}$ to $\boldsymbol{b}$ via m equally weighty unidimensional improvements means this: that there exists a life $\boldsymbol{a_1}$ which is a Pareto improvement on $\boldsymbol{a}$ and a Pareto worsening on $\boldsymbol{b}$, such that the unidimensional improvement $\boldsymbol{aa_1}$ is exactly as weighty as some improvement U, a life $\boldsymbol{a_2}$, which is a Pareto improvement on $\boldsymbol{a_1}$ and a Pareto worsening on $\boldsymbol{b}$, such that $\boldsymbol{a_1a_2}$ is exactly as weighty as U, ..., a life $\boldsymbol{a_{m-1}}$, which is a Pareto improvement on $\boldsymbol{a_{m-2}}$ and a Pareto worsening on $\boldsymbol{b}$, such that $\boldsymbol{a_{m-1}b}$ is exactly as weighty as U. If such m and n exist, and $m > n$, then the ab collection, taken as a whole, weighs more than the cd collection. Moreover, if $x = \frac{m}{n}$, the ab collection taken as a whole weighs x-times as much as the cd collection. The axioms above ensure that either such numbers m and n can be found, or else that as we "break down" the unidimensional improvements making up $\boldsymbol{ab}$ and $\boldsymbol{cd}$ into smaller and smaller improvements, all of equal weight except possibly for a last one of less weight, the ratio of the number of improvements of equal weight that approximately break down $\boldsymbol{ab}$ to the number of improvements of equal weight that approximately break down $\boldsymbol{cd}$ converges to some real number $\frac{m}{n}$. In this case, too, if $m > n$ we conclude that the ab collection weighs more than the cd collection; and if $x = \frac{m}{n}$, we conclude that the ab collection taken as a whole weighs x-times as much as the cd collection.[23]

This is how the account works in detail. It might be helpful to provide a statement of the account that is pithier than the one given over the last three paragraphs, but more informative than the one given on the first page of the paper. Here it is. According to the

[23] See Krantz et al. (1971, pp. 46-51, 142, 258-259); see also Nebel (2024, pp. 305-306) for a helpful exposition. Since I take conjoint measurement to offer a way to measure well-being differences, the account is closely related to difference measurement (see footnote 13 above). The latter is in turn a form of extensive measurement over "intervals" (Krantz et al. 1971, p. 143).

Pluralistic Tradeoffs Account, the well-being difference between lives a and b is greater than (resp. x-times as great as) the well-being difference between lives c and d if and only if, and, optionally, because, the collection of improvements from a to b is worth more than (resp. x-times as much as) that from c to d in contexts in which we need to make tradeoffs between improvements.

This is offered as a philosophical account of quantities of well-being. The claim is not that additive conjoint measurement provides the correct method for verifying certain empirical claims—say, about strength of preference.[24] Neither is it that conjoint measurement uncovers a conceptual or linguistic link between tradeoffs among goods and well-being quantities, which no competent thinker or speaker could possibly question (compare Williamson 2007). Rather, I am proposing the kind of biconditional with the optional inclusion of a claim of metaphysical dependence that is familiar from many areas of philosophy. The aim is to illuminate what stands on the left-hand side of the biconditional (quantities of well-being) in terms of what stands on the right-hand side (tradeoffs between improvements). We can call these *explanandum* and *explanans* respectively, though it is understood as a further optional claim that what stands on the left-hand side is metaphysically dependent on (say, grounded in or constituted by) what stands on the right-hand side.[25]

To provide a satisfactory defense of the **Pluralistic Tradeoffs Account**, we need to do two things.[26] One is to argue that the axioms on which it is based hold. We can call this the axiomatic claim. The

[24] We can now better appreciate a comparison between my account and the claim that conjoint measurement provides an interval scale for preference intensities (e.g., Suppes and Winet 1955, p. 260 and Bell and Raiffa 1988, pp. 388-395). In our setup, this latter claim requires us to defend the empirical assumption that people's preferences actually satisfy Normative Independence; my account requires us to defend the *normative* assumption that the betterness relation on lives satisfies Normative Independence. The two accounts come closer to one another if we restrict our attention to ideally rational self-interested agents. Even then, it is surely false that a necessary condition on *having* intensities of preferences is that one be ideally rational and self-interested. Therefore, the claim should be that ideally rational self-interested agents have intensities of preference that can be measured via conjoint measurement because such agents have intensities of preference that match quantities of well-being, and quantities of well-being can be measured via conjoint measurement.

[25] See Brenner et al. (2021) on varieties of metaphysical explanation. I remain neutral on the details; my point is simply that we should place the **Pluralistic Tradeoffs Account** on the same footing as, e.g., the claim that an act is wrong if and only if, and perhaps because, it has some given property, or that two time-slices belong to the same person if and only if, and perhaps because, some given relation holds between those two time-slices.

[26] I would like to thank Jake Nebel for suggesting that I tease apart these two components of my defense.

axioms entail that we can speak of *the* weight that any given collection of improvements gets when we need to make tradeoffs. The second thing to do is to argue that we are justified in interpreting this weight as a quantitative measure of well-being. We can call this the interpretative claim. When illustrating the account in Section 1.2, I attempted to back up the interpretative claim by relying on intuition. I now want to spell out in more detail why I think we should accept it. For the sake of this discussion, we provisionally accept the axioms—Section 1.5 will defend them.

1.4 A Defense of the Interpretation

Suppose that we have two different collections of improvements in various goods, with the former weighing more than (resp. x-times as much as) the latter in contexts in which we need to make tradeoffs. This means that whenever the former improvements are put to the test of being weighed against other things that are finally valuable for one's well-being, they matter more than (resp. x-times as much as) the latter improvements. If an improvement in well-being from a to b is greater than one from c to d, it better be that when the former is put to the test of being weighed against other things that are finally valuable for one's well-being, it is worth more than the latter. Well-being is normative: it is part of its role that greater quantities of it should be worth more when comparing lives in terms of which are better for the person living them.

But what about the converse implication? This is the claim that if a collection of improvements matters more than another whenever weighed against other things that are finally valuable for one's well-being, then it counts as a greater well-being difference. This seems quite intuitive, but it is certainly a more substantive claim. On reflection, however, it similarly allows us to make good sense of the roles that quantities of well-being play in normative theorizing.

To see the point, consider again the alleged counterexample to *leximin* mentioned in the introduction, used to illustrate why one might care about quantities of well-being. This time, however, we will describe the case somewhat differently. Suppose we can benefit the worst off individual

or the second worst off individual in society. Suppose that the improvement in well-being we could bestow on the second worst off individual is worth 100-times more than the improvement in well-being we could bestow on the worst off individual, when measured against other finally valuable goods (that is, in contexts in which we need to make tradeoffs). At present, the difference in well-being between the worst off and the second worst off individual consists in an improvement worth 1.01-times more than the improvement we could bestow on the worst off individual, when measured against other finally valuable goods. This seems about as close as we can get to describing the alleged counterexample to *leximin* without actually making any well-being difference comparison.[27] What this means is that the weight improvements in goods receive when we make tradeoffs matches the role that quantities of well-being play in normative theorizing. Someone looking to explain such quantitative comparisons in terms of more basic notions would have found an account that fits the bill. In short, I accept the interpretative claim because it vindicates the roles of well-being.

The next section argues for the axiomatic claim, with a special focus on the philosophically most interesting axiom, Normative Independence. Where a conclusive defense of a given axiom cannot be provided, it is shown that only small departures from the official account laid out in Section 1.3 are needed.

1.5 A Defense of the Axioms

I now turn to the axiomatic claim, discussing the axioms in the order in which I stated them.

[27] Again, supplementing the point of footnote 4, those wishing to deny interpersonal comparisons should redescribe the case as one concerning leximin as a decision theory under ignorance.

1.5.1 Normative Independence

Many philosophers have been skeptical of Normative Independence.[28] For example, Susan Hurley writes: "[...] the relationship between many different goods is *not* such as to meet this condition."[29] And John Broome (1991b, pp. 25-26) says: "[...] I do not see how to achieve any plausible separability in the dimension of sorts of goods"—"separability" being another name for Normative Independence. He continues: "I doubt that one sort could be judged independently from others" (p. 26). Hilary Greaves (2017, p. 202) mentions the idea that quantities of well-being might be constructed by adopting Normative Independence between "[...] 'consumption' along a number of dimensions (e.g. health, pollution level, house size, education level, food quality) [...]" but notes that "[t]he challenge for this approach is finding additional interpretations under which the separability condition really is plausible; this is not the case in the example just suggested."

A first point to note is that the "goods" these authors consider are in most cases merely instrumentally valuable things. This is clear in Greaves' example; Broome mentions in passing fame and comfort; Hurley discusses career success and family flourishing. I fully agree that these items are not normatively independent of one another, but want to suggest that this is so precisely because some of them are not finally valuable, real objective goods.[30] A simple illustration might help convey the idea: when I have a small house, a given improvement in education may matter more than one in fame. But if I have a large house, perhaps the opposite is true, since if I'm famous I can host huge parties and have a lot of fun. Notice, however, that I am here explaining why I make tradeoffs between education and fame differently depending on house size by reference to

[28] Of course, G. E. Moore's (1903) idea of "organic unities" is relevant here. My view is that the notion of organic unities can be understood in a number of ways and needs to be made precise—notably, with the help of measurement theory (see Carlson 2015). If it is made precise in terms of violations of Normative Independence, then my account obviously rules out organic unities between objective goods. On other ways of understanding the notion, my account is compatible with organic unities arising between particular instantiations of certain objective goods, as the discussion below will illustrate (roughly, the value that certain instantiations of objective goods generate when paired together is counted as an extra objective good, mirroring Moore's (1903, pp. 214-215) retention of a principle of additivity). Interactions between *degrees* of good are more problematic—see Section 1.6.

[29] Hurley (1989, p. 71, emphasis in original). To be precise, Hurley is here talking about what Keeney and Raiffa (1993, p. 91) call the corresponding tradeoffs condition, and what Krantz et al. (1971, p. 252) call the Reidemeister condition; this difference does not matter for our purposes.

[30] Compare Zimmerman (1983), Crisp (2000, pp. 37-40).

some extra good with respect to which house size and fame are instrumentally valuable—namely, the pleasure of having parties in a large house when I'm famous. Incidentally, this search for an explanation of what is merely instrumentally valuable should not be deemed in conflict with pluralism about well-being, which is at the core of the account: the value of friendship, unlike that of fame and house size, is arguably not reducible to the value of pleasure.

The foregoing suggests the conjecture that, when we lay out an exhaustive list of the finally valuable objective goods, they will turn out to be normatively independent of one another.[31] I now want to give a fully general argument to this effect, which I call the Explanatory-Link argument for Normative Independence.[32] This argument has three premises: the first connects improvements in objective goods to reasons for preference, the second connects reasons for preference to what we should prefer, and the third connects what we should prefer to what is better. Reasons are qualified to be normative and prudential; what you should prefer is qualified to be what you should prefer from the perspective of self-interest; and what is better is qualified to be what is better for well-being. I first state the three premises and then show with an example how they entail Normative Independence.

[31]Cf. Weirich (2001), who defends the idea that utility (understood as a measure of degrees of desire) is additive across what he calls "basic intrinsic" desires and aversions (pp. 48-49) by showing that this approach "[...] neither omits nor double-counts any relevant consideration" (p. 71). Leaving aside the distinction between desire intensities and quantities of well-being (for which see footnote 24), the most important contrast worth highlighting here is that Weirich assumes a quantitative notion of desire intensity, while my aim is to *derive* quantities of well-being.

[32]Somewhat ironically, my argument is structurally identical to Broome's (1991b) own argument for the sure-thing principle (pp. 95-96). Both Normative Independence and the sure-thing principle are, after all, structurally identical strong separability axioms. As we have seen above, however, Broome is skeptical of Normative Independence; my argument and Broome's do not stand or fall together. See Nebel (2023, pp. 1093-94) for a related line of reasoning in support of the independence axiom of extensive measurement for concatenation of "welfare-basic propositions via conjunction" (p. 1091). The Explanatory-Link argument can be understood as providing a *justification* for redescription strategies to salvage strong separability principles like Normative Independence, as discussed in Bell and Raiffa (1988), Hurley (1989), Oddie (2001), Greaves (2017). In particular, it backs up Oddie's proposal that we use additivity as a "regulative ideal" (p. 331) or "heuristic tool" (p. 330) "[...] in the search for intrinsic goods" (ibid.).

Premises of the Explanatory-Link Argument.

Improvements-Link: The only reasons for preferring a life $\boldsymbol{a}$ to another $\boldsymbol{b}$ are improvements from $\boldsymbol{a}$ to $\boldsymbol{b}$ along finally valuable dimensions of well-being.

Reasons-Link: If you have the same reasons for preferring life $\boldsymbol{a}$ to $\boldsymbol{b}$ as you have for preferring life $\boldsymbol{c}$ to $\boldsymbol{d}$, and you have the same reasons for preferring $\boldsymbol{b}$ to $\boldsymbol{a}$ as you have for preferring $\boldsymbol{d}$ to $\boldsymbol{c}$, then you should prefer $\boldsymbol{a}$ to $\boldsymbol{b}$ iff you should prefer $\boldsymbol{c}$ to $\boldsymbol{d}$.

Preference-Link: You should prefer $\boldsymbol{a}$ to $\boldsymbol{b}$ iff $\boldsymbol{a}$ is better than $\boldsymbol{b}$.

Consider now the following example, again assuming for simplicity that there are only three objective goods: pleasure, knowledge, and friendship. Lives $\boldsymbol{a}$ and $\boldsymbol{b}$ in the diagram below are alike with respect to friendship, with $\boldsymbol{a}$ doing better in terms of pleasure, and $\boldsymbol{b}$ doing better in terms of knowledge. Lives $\boldsymbol{c}$ and $\boldsymbol{d}$ are exactly like $\boldsymbol{a}$ and $\boldsymbol{b}$ respectively, except that they are both at a lower, and equal, degree of friendship.

$\boldsymbol{a}$: (🙂, 🎓, 🎉) $\succ$ $\boldsymbol{b}$: (😐, 🎓🎓, 🎉)

$\boldsymbol{c}$: (🙂, 🎓, 🕴) _ $\boldsymbol{d}$: (😐, 🎓🎓, 🕴)

Figure 1.3

If Normative Independence holds, then $\boldsymbol{c} \succ \boldsymbol{d}$. A comparison between $\boldsymbol{a}$ and $\boldsymbol{b}$ involves a tradeoff between pleasure and knowledge, and the fact that $\boldsymbol{a} \succ \boldsymbol{b}$ means that pleasure wins out. If Normative Independence holds, how tradeoffs are made between pleasure and knowledge is independent of what degree of friendship we keep fixed. Since the comparison between $\boldsymbol{c}$ and $\boldsymbol{d}$ involves that very same tradeoff, given Normative Independence $\boldsymbol{c} \succ \boldsymbol{d}$.

We now derive this conclusion from our three premises. Note that the only improvement from $\boldsymbol{b}$ to $\boldsymbol{a}$ is in pleasure, and it is the very same improvement from $\boldsymbol{d}$ to $\boldsymbol{c}$. So, by Improvements-Link, you have the very same reasons for preferring $\boldsymbol{a}$ to $\boldsymbol{b}$ as you have for preferring $\boldsymbol{c}$ to $\boldsymbol{d}$. The only improvement from $\boldsymbol{a}$ to $\boldsymbol{b}$ is in knowledge, and it is the very same improvement from $\boldsymbol{c}$ to $\boldsymbol{d}$. So, by

Improvements-Link, you have the very same reasons for preferring $\boldsymbol{b}$ to $\boldsymbol{a}$ as you have for preferring $\boldsymbol{d}$ to $\boldsymbol{c}$. By Reasons-Link, then, you should prefer $\boldsymbol{a}$ to $\boldsymbol{b}$ if and only if you should prefer $\boldsymbol{c}$ to $\boldsymbol{d}$. Since $\boldsymbol{a} \succ \boldsymbol{b}$, by Preference-Link you should prefer $\boldsymbol{a}$ to $\boldsymbol{b}$. So, you should prefer $\boldsymbol{c}$ to $\boldsymbol{d}$. By another application of Preference-Link, $\boldsymbol{c} \succ \boldsymbol{d}$.

I used an example to illustrate the point, but the argument applies in full generality. Any test-case for Normative Independence will involve some affected goods and some non-affected goods. We will have two lives, $\boldsymbol{a}^*$ and $\boldsymbol{b}^*$, with $\boldsymbol{a}^*$ doing better than $\boldsymbol{b}^*$ in some affected goods, and $\boldsymbol{b}^*$ doing better than $\boldsymbol{a}^*$ in some others. We will then have a verdict, say, $\boldsymbol{a}^* \succ \boldsymbol{b}^*$. Next, we come up with two other lives $\boldsymbol{c}^*$ and $\boldsymbol{d}^*$, exactly like $\boldsymbol{a}^*$ and $\boldsymbol{b}^*$ respectively except that the non-affected goods are kept fixed at a different level. Then Normative Independence requires that $\boldsymbol{c}^* \succ \boldsymbol{d}^*$. We use (i) Improvement-Link to determine that the $\boldsymbol{a}^*$ vs. $\boldsymbol{b}^*$ comparison is exactly like the $\boldsymbol{c}^*$ vs. $\boldsymbol{d}^*$ comparison with respect to reasons; (ii) Reasons-Link to determine that the $\boldsymbol{a}^*$ vs. $\boldsymbol{b}^*$ comparison is exactly like the $\boldsymbol{c}^*$ vs. $\boldsymbol{d}^*$ with respect to what you should prefer; and (iii) Preference-Link to determine that the $\boldsymbol{a}^*$ vs. $\boldsymbol{b}^*$ comparison is exactly like the $\boldsymbol{c}^*$ vs. $\boldsymbol{d}^*$ comparison with respect to which lives are better for well-being. Since $\boldsymbol{a}^* \succ \boldsymbol{b}^*$, we get that $\boldsymbol{c}^* \succ \boldsymbol{d}^*$.

This is the Explanatory-Link argument. All of its premises are, I believe, *prima facie* very intuitive. But they also receive support from general philosophical approaches that in no way prefigure a commitment to Normative Independence. Improvements-Link is supported by a picture of reasons as grounded in values, together with a picture of value that reduces goodness to betterness (Wedgwood 2022, p. 131). Reasons-Link seems to fit well with a picture of rational preference as reasons-responsiveness (Kiesewetter 2017; Lord 2018). Preference-Link hardly needs any defense given the qualification that we are talking about self-interested preference and betterness for well-being. Yet, since these premises entail Normative Independence, a staunch objector will want to reject at least one of them.[33]

[33]Picking up on footnote 32: Broome (1991b) thinks his argument for the sure-thing principle begs the question, since it relies on the premise that "[...] every reason there is for preferring one alternative to another must derive from what happens in some state of nature, taken by itself" (p. 96), which he says is "[...] the essence of the sure-thing principle" (ibid.). The analogous premise in my argument would be Improvements-Link, but it seems to me a stretch

Preference-Link cannot plausibly be rejected. What about denying Improvements-Link? The idea would have to be that there can be reasons to prefer a life $\boldsymbol{a}$ over another, $\boldsymbol{b}$, that are not finally valuable ways in which $\boldsymbol{a}$ does better than $\boldsymbol{b}$ for well-being. But if something is not a way in which $\boldsymbol{a}$ does better than $\boldsymbol{b}$, how could it give us any reason to prefer $\boldsymbol{a}$ over $\boldsymbol{b}$? Consider the following schema:

> One reason to prefer life $\boldsymbol{a}$ over $\boldsymbol{b}$ is that _____

Minimally, whatever fills in the blank will have to be something *about* $\boldsymbol{a}$ and $\boldsymbol{b}$. How could it not be a respect in which $\boldsymbol{a}$ is better than $\boldsymbol{b}$? The onus is on the opponent of Improvements-Link to answer this question; I cannot see how the challenge could be met.

This leaves us with Reasons-Link. An opponent might say: although you have the *same* reasons in the $\boldsymbol{a}$ vs. $\boldsymbol{b}$ situation as you do in the $\boldsymbol{c}$ vs. $\boldsymbol{d}$ situation, perhaps the differing levels of the non-affected goods in the two situations alter the strength of these reasons—they act as what get called *modifiers*, conditions that are not reasons but affect the strength of reasons (Dancy 2004; Bader 2016). Consider again our earlier example. If Reasons-Link fails in this way, we might have $\boldsymbol{d} \succ \boldsymbol{c}$.

$$\boldsymbol{a} : (🙂, 🎓, 🎉) \succ \boldsymbol{b} : (🙂, 🕶️🕶️, 🎉)$$

$$\boldsymbol{c} : (🙂, 🎓, 🧍) \prec \boldsymbol{d} : (🙂, 🕶️🕶️, 🧍)$$

Figure 1.4

What explains this verdict, the objection goes, is that the improvement in pleasure from $\boldsymbol{b}$ to $\boldsymbol{a}$ weighs more than the improvement in pleasure from $\boldsymbol{d}$ to $\boldsymbol{c}$, even if they are the very same improvement in pleasure, and hence must provide the very same reason.

---

to say that Improvements-Link is the essence of Normative Independence. In fact, the objector's best bet may be to challenge Reasons-Link, as we will see below. Charges of "begging the question" are notoriously hard to adjudicate; at the very least, the Explanatory-Link argument offers a helpful framework for assessing Normative Independence that connects this axiom to commitments about reasons, preference, and value—as I see it, this makes it an argument worth thinking about.

We must be careful here. The improvements from $\boldsymbol{b}$ to $\boldsymbol{a}$ and from $\boldsymbol{d}$ to $\boldsymbol{c}$ are in degrees of goods, and involve abstract lives. The story cannot be, for example, that the very same pleasure weighs more when enjoyed with friends, as opposed to when enjoyed alone. For, in at least some of the concrete lives that instantiate the given degrees of friendship and pleasure, the way your life goes with respect to pleasure is virtually unrelated to the way it goes with respect to friendship. In one such life, for example, you get 40 years of pleasure in the experience machine, and then go through some real-life hurdles with newfound friends for 40 more years. It seems utterly implausible that the level of friendship in this life intensifies the weight of pleasure, which would have to be the case for the objector's point to apply to degrees of goods. In general, there seems to be no plausible rationale for thinking that different *degrees* of friendship act as modifiers for how much degrees of pleasure matter. Similar points apply to other goods.

In light of this observation, the objector might reply: so much the worse for abstract lives! They might become convinced that the problem lies, in the first instance, with a presupposition of the Explanatory-Link argument: the Pareto axiom, and specifically its implication—which we can call Pareto Indifference—that two concrete lives doing equally well in all objective goods are equally good (compare Hedden and Nebel 2024, p. 288). For suppose we run with the idea that certain *instantiations* of degrees of friendship can act as modifiers for instantiations of degrees of pleasure. Then there may be four concrete lives a, b, c, d instantiating the degrees of objective goods displayed by $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}$, and such that $a \succ b$ but $d \succ c$, with a rejection of versions of Normative Independence and Reasons-Link for concrete lives. For some other a^*, b^* instantiating $\boldsymbol{a}, \boldsymbol{b}$, respectively, but such that the instantiation of the degree of friendship in a^* does not similarly intensify the weight of the improvement in pleasure from b^* to a^* (because, say, the pleasure is unrelated to the friendship), we will have $b^* \succ a^*$—reflecting the point that the very same improvement in degree of pleasure weighs more when enjoyed with friends than when caused by an experience machine. Assuming $b \sim b^*$, this entails the negation of $a \sim a^*$, violating Pareto Indifference.

The problem now is that the objector faces a difficult choice: if they are to retain Improvements-

Link for concrete lives, they need to say that a is better than a^* even if there is no reason to prefer a to a^*. This, however, amounts to a more radical violation of Reasons-Link for concrete lives, one that is not explained by modifiers, since for there to be a modifier there needs to be a reason. Alternatively, they can say that there are reasons to prefer a to a^* that are not ways in which a is better than a^*, thereby rejecting Improvements-Link, as well as Reasons-Link, for concrete lives. As we saw above, however, denying Improvements-Link seems rather unappealing. Thus, on the way to establishing Normative Independence, the premises of the Explanatory-Link argument secure Pareto Indifference, too.[34] Those who disagree should still find the argument of some interest: it is clear that Pareto Indifference is much less contentious than Normative Independence, and the argument's lesson is at the very least that to reject the latter, we must actually reject the former.

We now turn to a second possible line of attack on Reasons-Link. According to this view, what modifies the strength of reasons is not instantiations of other goods, but rather facts about certain "global properties" of a life (compare Buchak 2013, pp. 38-47, on global properties of gambles). One could think, for example, that improvements along the dimension on which one does *worst* should be given extra weight. It might then be that, although you have the same reasons in $\boldsymbol{a}$ vs. $\boldsymbol{b}$ as you do in $\boldsymbol{c}$ vs. $\boldsymbol{d}$, the given pleasure improvement amounts to an improvement in the worst dimension in $\boldsymbol{a}$ vs. $\boldsymbol{b}$ but not in $\boldsymbol{c}$ vs. $\boldsymbol{d}$. For this reason it gets less consideration in the latter case, resulting in $\boldsymbol{d} \succ \boldsymbol{c}$ and a violation of Normative Independence.

This idea, however, requires us to be able to rank individual goods in terms of the level of well-

[34] Again, this can be seen as a justification of redescription strategies for Pareto Indifference in our context—see Hedden and Muñoz (2024), as well as Hedden and Nebel (2024, pp. 288-290). Hedden and Muñoz formulate their Pareto principles in a way that does not entail Pareto Indifference; however, they extend their redescription approach to argue for "Dimensionalism," which is also Hedden and Nebel's focus in the pages just mentioned. Dimensionalism is equivalent to Pareto Indifference in our single-profile setup (i.e., in which the orderings $\succeq_i$ are fixed). A common deflationary treatment of the kind of context-sensitivity exemplified by modifiers is indeed that it results from a non-exhaustive specification of what reasons one has in each case (Hooker 2000; Raz 2000). Bader (2016) aims to rehabilitate context-sensitivity, and *inter alia* the distinction between reasons and modifiers. But his argument for this latter point rests on the claim that there is such a thing as the strength of the "unmodified reason," which can be determined if one "[...] restrict[s] the evaluation to the ground of the reason, assessing *to what extent* something generates reasons in virtue of having such a ground" (p. 40, emphasis added). In our setting, this amounts to postulating an unmodified weight for each improvement in any given good, which is illicit given our methodological assumption that we need to earn our right to talk about quantities of well-being.

being they instantiate. And I deny that this makes any sense. Well-being is what's good *for a person*; a level of well-being is an equivalence class of equally good lives for a person. Simply put: a good, like friendship, is not a person; it therefore cannot have a level of well-being. It might be suggested that we can derive levels of well-being for goods from ordinal comparisons, in the following way. Our aim, for example, might be to establish whether some degree of knowledge is at the same well-being level as some degree of pleasure. We look at a life, $\boldsymbol{e}$, in which all goods but knowledge are at their lowest level, and knowledge is at the degree of interest. We then look at a life, $\boldsymbol{f}$, in which all goods but pleasure are at their lowest level, and pleasure is at the degree of interest. If $\boldsymbol{e} \succ \boldsymbol{f}$, it might seem reasonable to conclude that the well-being level of knowledge in $\boldsymbol{e}$ is higher than the well-being level of pleasure in $\boldsymbol{f}$. But far from deriving well-being levels for goods from ordinal comparisons, this proposal in fact assumes that the lowest degrees of each good are at the same well-being level. For if this were not the case, it is unclear why $\boldsymbol{e} \succ \boldsymbol{f}$ should allow us to conclude anything about the well-being level of knowledge in $\boldsymbol{e}$ and of pleasure in $\boldsymbol{f}$.[35]

It should finally be remarked that even if we could somehow make sense of levels of well-being for each good, the current motivation for rejecting Normative Independence would still support a "comonotonic" restriction of the axiom. On this weakening, Normative Independence holds within sets of lives that rank goods in the same way. A comonotonic version of Normative Independence can still be used to obtain additive representations (see Wakker 1991; see also Schmeidler 1989, Wakker 1989, Köbberling and Wakker 2003, Buchak 2013).[36] Thus, the general claim that quantities of well-being arise from tradeoffs between objective goods would only need to be modified: if violations of Normative Independence are of this rank-based kind, quantities of well-being arise from tradeoffs between objective goods in comonotonic sets of lives.[37]

[35] Another difficulty for this approach is that there may not be a lowest level for some goods, if some goods are unbounded.

[36] To be precise, there is a distinction between imposing Normative Independence on each comonotonic set, and what may naturally be called Comonotonic Normative Independence, which would additionally impose invariance on how tradeoffs between well-being levels are made within different comonotonic sets—see Köbberling and Wakker (2003, p. 400).

[37] I thank Jake Nebel and Lara Buchak for pointing out this alternative route to me. My argument against the idea of well-being levels for goods will anyway be relevant in Section 1.6—see footnote 44.

1.5.2 Restricted Solvability

Roughly, Restricted Solvability says the following: as long as we are sufficiently far from the best or worst levels of any good (if there are such maximum and minimum levels), we will be able to exactly compensate losses in some goods with improvements in other goods. Consider a life which is nowhere near the best or worst level of pleasure and knowledge (again, if such levels exist). And suppose we entertain a loss in pleasure relative to it. According to Restricted Solvability, we will be able to find an improvement in knowledge such that getting the improvement in knowledge exactly compensates the loss in pleasure.

Since we are evaluating all possible lives in terms of their well-being, our search for exact compensations of this kind is constrained by nothing but what holds of *metaphysical* necessity. This arguably gives us as much freedom as we need. Suppose we are looking to exactly compensate a loss in pleasure with an improvement in knowledge. Learning all biology up to the latest research will more than compensate, but reading a book about giraffes will not be enough. What about reading a book about about giraffes and one about insects? If this is still not enough, what about learning all biology studied at an undergraduate level? If this is too much, what about learning all biology studied at an undergraduate level but forgetting it 5 years before one dies? What about forgetting it 4 years, 341 days, and 31 seconds before one dies? Given this embarrassment of riches, the prospect for an exact compensation seems promising. This is just an example—the point applies quite generally to any objective good.

I admit that the support this consideration lends to Restricted Solvability falls rather short of conclusive argument. It is therefore worth pointing out that axioms of "solvability" are non-necessary, or structural (Krantz et al. 1971, p. 23). In our context: it doesn't follow from the fact that well-being comes in quantities that Restricted Solvability is satisfied. In fact, in settings of conjoint measurement somewhat different from the one used here, additive representations have been obtained with solvability conditions imposed on some components only (Gonzales 1996, 2000, 2003). Since illuminating necessary and sufficient conditions for additive representations using

conjoint measurement are not known in the general case, a defense of the **Pluralistic Tradeoffs Account** is bound to overshoot. Skeptics should bear in mind that a counterexample to Restricted Solvability is not necessarily a counterexample to the existence of an additive representation for well-being.

Even if failures of Restricted Solvability prevented the possibility of an additive representation, not much would be lost: we would just need to make room for a small amount of incompleteness in our theory of quantities of well-being.[38] Assume for the sake of illustration that Restricted Solvability does not hold because some good i happens to have a more discriminating structure than the other goods. We can still get many quantitative well-being comparisons via the following method. We take the ordering on good i to be more coarse-grained than it actually is. We "weed out" of our model some degrees of good i in a way that makes the structure of the resulting ordering on good i match the structure of the other goods. Suppose we select four concrete lives, a, b, c, d, with b better than a, and d better than c. We want to know whether the difference in well-being between a and b is greater than that between c and d. Unfortunately, the improvement in good i from a to b is not one that our simplified model entertains. Even so, our model will entertain two approximately equal improvements, one slightly worse and one slightly better. If, no matter which one we pick, we get that the ab collection is worth more than the cd collection, then we can conclude that the well-being difference between a and b is greater than that between c and d. If our choice makes a difference, then our account is not able to compare the well-being difference between a and b with that between c and d. Proceeding in a similar fashion, the strategy can also be implemented to reach many verdicts concerning whether one difference in well-being is at least or at most x-times larger than another.

1.5.3 Archimedeanism

Archimedeanism says that all objective goods can be traded off, in the sense that none is infinitely more valuable than the others. If a good is infinitely more valuable than another, or lexically supe-

[38] Contrast what follows here with the large amount of semantic indeterminacy entertained by Greaves (2017).

rior, then any increase in the former good is worth any decrease in the latter.[39] If a good is lexically superior to another, then there are simply no tradeoffs to be made between the two.

At least when we look at typical goods that have been defended as featuring on the objective list of well-being, it seems plausible to think that tradeoffs *can* be made. A life of knowledge without pleasure is dull, suggesting pleasure counts for something when pitted against knowledge. A life of pleasure without knowledge is crass, suggesting knowledge counts for something when pitted against pleasure. One of achievement without friendship self-absorbed, suggesting friendship counts for something when pitted against achievement. And so on. There are, it seems, no hierarchies in goods. Each increase in one good counts for something relative to sacrifices in another good. Extra support for this premise is lent by the fact that a version of this axiom appears in orthodox decision theory, often under the name of "continuity"—in that context, it guarantees the possibility of tradeoffs in probability, so that no outcome is such that any positive probability of it whatsoever swamps all other considerations (see, e.g., Kreps 1988, pp. 44-46; Köbberling and Wakker 2003, pp. 398-99).[40]

Of course, a number of disparate philosophical and religious traditions do entertain the idea that some goods are lexically superior to others. Salvation, however understood, should presumably count as infinitely more valuable than any earthly good. For a secular approach, consider Mill's theory of higher and lower pleasures, which can be seen as positing a relationship of lexical priority between two qualitatively different kinds of enjoyment (see Nebel 2022b). If any of these theories is correct, the account proposed here will no doubt need to be modified, but in most cases not abandoned. While Archimedeanism is necessary for a real-valued additive representation, it is

[39]The literature often distinguishes between strong and weak superiority, with strong superiority defined as in the main text, and weak superiority defined as follows: a sufficiently large increase in the former good is worth any decrease in the latter good (see Nebel 2022b, p. 209). The framework of conjoint measurement only rules out strong superiority, since no axiom of this framework forbids the following scenario: some good is bounded, some other good is unbounded, and a sufficiently large increase in the unbounded good is worth more than the largest possible increase in the bounded good.

[40]Therefore, proponents of the popular vNM method for constructing quantities of well-being mentioned in the introduction are in no position to object to Archimedeanism in our context.

not necessary for making a whole range of quantitative well-being comparisons.[41]

Begin with the case of a single superior good. In that event, the account of this paper should be understood as one of quantities of well-being for the inferior goods in the first place. If we assume that the superior good is "on-off"—either you have it or you don't—the account can then become one of well-being quantities *tout court*. Suppose we select four concrete lives, a, b, c, d, with b better than a, and d better than c. There are four options: (i) all lives have or lack the superior good; (ii) only b and d have it; (iii) only b has it; (iv) only d has it. If (i) or (ii), we follow the procedure of Section 1.3 to determine whether the improvement in inferior goods from a to b is greater than that from c to d; if that is the case, the well-being improvement from a to b is greater than that from c to d. If (i), moreover, it seems right to say that when the improvement in inferior goods from a to b is x-times as large as that from c to d, then the well-being improvement from a to b is x-times as great as that from c to d. However, it seems wrong to say that in case (ii). In this case, the difference between the two improvements in well-being is infinitesimally small, and so the account delivers no precise ratio between them. If (iii) or (iv), the improvement in well-being from inferior to superior goods is greater than that in merely inferior goods. It is, in fact, infinitely greater, and so the account delivers no precise ratio between the two improvements in well-being.

Suppose, next, that there are two tiers of objective goods, with each tier hosting multiple goods—for example, the higher pleasures and knowledge belong to the lexically superior tier, and the lower pleasures and avoidance of bodily pains belong to the lower tier. Then versions of the **Pluralistic Tradeoffs Account** can be developed for each tier, and stitched together to arrive at an overall theory of quantities of well-being.[42] Suppose we select four concrete lives, a, b, c, d, with b better than a, and d better than c. If the improvement from a to b in the higher goods is greater than that from c to d, then the well-being improvement from a to b is greater. If the improvement from

[41] See Carlson (2007) for a general model of non-Archimedean additivity.

[42] While my official account of Section 1.3 requires at least three goods, variations of that account can be given for two goods only (see Krantz. et al 1971, p. 257, THEOREM 2). Normative Independence then needs to be supplemented with a higher-order "cancellation axiom" (which is entailed by Normative Independence and the other axioms in the case of three or more goods).

a to b in the higher goods is equal to that from c to d, we look at the lower goods to break the tie. If going from a to b improves things x-times as much as going from c to d with respect to the higher goods, and the two improvements are equal with respect to the lower goods, then the well-being improvement from a to b is x-times as great as that from c to d. However, if the two improvements are different with respect to the lower goods, the account delivers no ratio between the two improvements in well-being, though it seems right to say that the well-being improvement from a to b is approximately x-times as great as that from c to d.

More details could be given on how to develop or limit the scope of the **Pluralistic Tradeoffs Account** in contexts in which the Archimedean axiom does not hold. My brief remarks are only supposed to show that, while rejecting this axiom requires a departure from the official view of Section 1.3, such a departure need not involve abandoning the core of that account.

1.6 Metaphysical Independence

To recap, the three axioms of the **Pluralistic Tradeoffs Account** can defended; where they cannot be established conclusively, it can be shown that only minor departures from the official account are called for, should the relevant axiom fail. Is our job done? Not quite. I need to return to an assumption I made in Section 1.3, Metaphysical Independence.

On inspection, it seems a live possibility that Metaphysical Independence could fail. For example, some people think that knowledge is an achievement (Greco 2012). If we accept this view and assume that knowledge and achievement are objective goods, no possible concrete life will instantiate both a high degree of knowledge and a low degree of achievement. Or consider a view on which the pinnacles of friendship can only be achieved by very knowledgeable people. On this view, there is no possible concrete life that instantiates both a high degree of friendship and low degree of knowledge. Let us call these "high-low" violations of Metaphysical Independence, in that they forbid combinations of "high" levels of some good with "low" levels of another.

These violations of Metaphysical Independence are unlikely to change the picture presented here. While the representation theorem I appealed to (by Krantz et al. 1971) assumes a Cartesian product structure, this assumption is not necessary (Krantz et al. 1971, pp. 275-276), and various results confirm that conjoint measurement can be performed on a subset of a Cartesian product, to obtain additive representations (Jaffray 1974; Segal 1992; Chateauneuf and Wakker 1993). Each of these results imposes some amount of structure on the subset of interest, and they are proved in slightly different settings (Segal 1992 and Chateauneuf and Wakker 1993, for example, make assumptions of topological connectedness). But the general lesson they point to, I believe, is that violations of Metaphysical Independence of the "high-low" kind are compatible with the **Pluralistic Tradeoffs Account**.

There is however a much more radical way in which Metaphysical Independence could fail. Suppose it is valuable, or better, to have certain particular combinations of degrees of objective goods: for example, any life with *these* degrees of knowledge, pleasure, and so on, is, in a finally valuable respect, better than any life with *those* degrees of knowledge, pleasure, and so on, independently of the ways in which it might be better because of improvements in knowledge, pleasure, and so on. Then, we need to make space in our objective list of well-being for a "pattern good" (Broome 2004, p. 44) that concerns what particular degrees of the other goods one's life instantiates. In the extreme case, this view leads to the conclusion that any degree of any objective good appears in exactly one abstract life: any other combination of degrees of goods in which it appears is not instantiated by any concrete life. This would invalidate the **Pluralistic Tradeoffs Account** because any tradeoff between goods would turn out to be metaphysically impossible. Recall that, in order to make tradeoffs between two goods, we need to be able to vary degrees of those goods, while keeping fixed degrees of all other goods. And the view under discussion denies we can do this.[43]

How could such a view be motivated? The claim is that the combination of *this* degree of knowl-

[43]Returning to the issue of Moore's (1903) "organic unities," discussed in footnote 28: the view under consideration would be an example of a kind of organic unity incompatible with my account.

edge, however instantiated, together with *that* degree of friendship, however instantiated, and so on, brings about some extra value. How could mere combinations of degrees of objective goods matter in this way? The most intuitive motivation I can think of is this: equality or balance between different non-pattern goods matters for well-being, and different combinations of degrees of non-pattern goods are more or less balanced.

Equality may be valuable in other contexts. For example, it may be morally valuable for well-being to be distributed equally across a population of individuals. But it is hard to make sense of the idea that equality between objective goods might be valuable, for the technical reason that levels of goods on the objective list of well-being are plausibly non-comparable (see Hedden and Nebel 2024, p. 279). To say that there is more pleasure *here* than friendship *there* makes no more sense than to say that three meters are greater than five kilograms. A balanced life would be one in which one has equal amounts of each good; but such a notion makes no sense, on inspection.[44]

It can be replied that perhaps we can make rough or imprecise level comparisons, as evidenced by our use of expressions such as "a very knowledgeable life," or "a life with little pleasure." The idea would then be that we can use such expressions to pick out lives we judge to be more balanced, and lives we judge to be less balanced. On this modification of the view, there will surely be very many lives that are neither better nor worse than one another with respect to balance; but we will nonetheless be able to identify some clear cases of more and less balanced lives. The ensuing violation of Metaphysical Independence would be less severe, but the worry might remain that it will prevent the **Pluralistic Tradeoffs Account** from working properly.

Here is, then, a more direct response to this problem. If some objective goods come in quantities of their own—as is plausible, for example, in the case of pleasure—we can perfectly well accommodate the intuition that motivates the current proposal about the value of balance without

[44] A different way to understand equality or balance would be to say that a balanced life is one in which all goods are at equal well-being levels. I do not discuss this option in the main text since we already rejected the thought that we can compare well-being levels for goods in Section 1.5.1, when discussing rank-based objections to Normative Independence.

thinking of balance as a finally valuable dimension of well-being, and therefore without having to postulate widespread violations of Metaphysical Independence. To illustrate, suppose that indeed there are quantities of pleasure. Then, our betterness ordering on lives may very well encode the idea that increasing pleasure in a life carries diminishing marginal contributions to that life's well-being: keeping fixed a sacrifice in some other good, the less pleasure one has the smaller the improvement in pleasure will have to be for it to exactly counterbalance that given sacrifice. In short, more and more pleasure matters less and less for well-being.

While there may not be quantities of some other goods, we will be able to tell a similar story for each good, so that having more and more of any good matters less and less for well-being. How? We are spotting to the objector the idea that we can make rough level comparisons. Pick two levels of knowledge $k1$ and $k2$, and some levels of pleasure $p1$ and $p2$ they are approximately as great as. Next, pick two higher levels of pleasure $p3$ and $p4$ such that the $p3$-$p4$ difference in pleasure equals in size the $p1$-$p2$ difference, and such that there are levels of knowledge $k3$ and $k4$ which $p3$ and $p4$ are respectively approximately as great as. Then, our betterness ordering on lives may very well encode the idea that increasing knowledge in a life carries diminishing marginal contributions to that life's well-being: if a sacrifice in another good exactly makes up for an increase in knowledge from $k1$ to $k2$, a smaller sacrifice of that good will be needed to make up for an improvement from $k3$ to $k4$. If our betterness relation looks like this, it will vindicate verdicts such as that, instead of adding two units of pleasure, or improve knowledge "by a lot," it would be better to add a unit of pleasure together with "a more moderate" improvement in knowledge. Intuitively more balanced lives will be better for well-being, with no violations of Metaphysical Independence. Balance, on this view, is not an additional objective good but rather a feature of the betterness relation.

1.7 Conclusion

This paper took on board an important methodological remark of Broome's *Weighing Goods*:

> I cannot see what use we can have for the notion of quantities of good except when we weigh differences in good in comparing alternatives. So it is in weighing up differences that we can expect the notion to get its meaning. (Broome 1991b, p. 215)

While the title may suggest otherwise, Broome is skeptical in the book that the relevant weighing is to be done between *kinds* of goods. In this paper, I hope to have provided a suitably comprehensive defense of the idea that, to the contrary, this is exactly how quantities of personal good, or well-being, arise.

Appendix

Here is a formal statement of the account. We begin with a statement of the axioms, following Krantz et al. (1971, page references below are to this work).

For every subset of goods $M \subset N$, we can divide any $\boldsymbol{a} \in \mathcal{L}$ into two components, $\boldsymbol{a_M} = (\boldsymbol{a_i})_{i \in M} \in \mathcal{L}_M$ and $\boldsymbol{a_{N \setminus M}} = (\boldsymbol{a_i})_{i \in N \setminus M} \in \mathcal{L}_{N \setminus M}$, so that $\boldsymbol{a}$ can be written as $(\boldsymbol{a_M}, \boldsymbol{a_{N \setminus M}})$. We say that M is separable under $\succeq$ if and only if for all $\boldsymbol{e}, \boldsymbol{f} \in \mathcal{L}_M$ and $\boldsymbol{g}, \boldsymbol{h} \in \mathcal{L}_{N \setminus M}$, $(\boldsymbol{e}, \boldsymbol{g}) \succeq (\boldsymbol{f}, \boldsymbol{g})$ if and only if $(\boldsymbol{e}, \boldsymbol{h}) \succeq (\boldsymbol{f}, \boldsymbol{h})$.

Normative Independence. All subsets of N are separable. (p. 301)

Restricted Solvability. For all $i \in N$, if we have $(\boldsymbol{b_1}, ..., \bar{\boldsymbol{b}}_{\boldsymbol{i}}, ..., \boldsymbol{b_n}) \succeq (\boldsymbol{a_1}, ..., \boldsymbol{a_i}, ..., \boldsymbol{a_n}) \succeq (\boldsymbol{b_1}, ..., \underline{\boldsymbol{b}}_{\boldsymbol{i}}, ..., \boldsymbol{b_n})$ then we can find $\boldsymbol{b_i} \in \mathcal{L}$ such that $(\boldsymbol{a_1}, ..., \boldsymbol{a_i}, ..., \boldsymbol{a_n}) \sim (\boldsymbol{b_1}, ..., \boldsymbol{b_i}, ..., \boldsymbol{b_n})$. (p. 301)

A standard sequence (p. 253) in good $i \in N$ is a set of degrees of good i, $\boldsymbol{a^1}, \boldsymbol{a^2}, ... \in \mathcal{L}_i$ that are equally spaced, in the sense that for some other good $j \in N$, there exist degrees $\boldsymbol{b^+} \succ_{\boldsymbol{j}} \boldsymbol{b}$, such that for all $\boldsymbol{a^i}, \boldsymbol{a^{i+1}}$ in the sequence, $(\boldsymbol{a^{i+1}}, \boldsymbol{b}) \sim (\boldsymbol{a^i}, \boldsymbol{b^+})$, omitting the constant component in $N \setminus \{i, j\}$. A standard sequence in good i is strictly bounded if there exist a degree of good i better than any of its members, and a degree of good i worse than any of its members.

Archimedeanism. "Every strictly bounded standard sequence is finite." (p. 256)

Given these axioms and the framework of Section 1.3, there exist real-valued functions ϕ_i on $\mathcal{L}_i$, for all $i \in N$, such that for all $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{L}$

$$(\boldsymbol{a_1}, ..., \boldsymbol{a_n}) \succeq (\boldsymbol{b_1}, ..., \boldsymbol{b_n}) \text{ iff } \phi_1(\boldsymbol{a_1}) + ... + \phi_n(\boldsymbol{a_n}) \geq \phi_1(\boldsymbol{b_1}) + ... + \phi_n(\boldsymbol{b_n}); \tag{1.1}$$

moreover, if another set of functions ϕ_i' is such as to satisfy (1.1), then $\phi_i' = \alpha\phi_i + \beta_i$, with $\alpha > 0$ and constant (THEOREM 13, p. 302). Note that the statements of these axioms would need to be modified slightly if, in line with Section 1.6, we were to relax the assumption of Metaphysical Independence, that $\mathcal{L} = \mathcal{L}_1 \times ... \times \mathcal{L}_n$.

Let $\phi = \phi_1 + ... + \phi_n$, with the set of ϕ_i satisfying (1.1). Consider four concrete lives $a, b, c, d \in L$. If $\phi(\boldsymbol{b}) - \phi(\boldsymbol{a}) \geq \phi(\boldsymbol{d}) - \phi(\boldsymbol{c})$, then the collection of improvements from a to b is worth at least as much as that from c to d, when tradeoffs need to be made. The account can be stated formally as follows:

> (**Pluralistic Tradeoffs Account**) The well-being difference between lives a and b is at least as great as (resp. x-times as great as) that between lives c and d if and only if, and, optionally, because, $\phi(\boldsymbol{b}) - \phi(\boldsymbol{a}) \geq \phi(\boldsymbol{d}) - \phi(\boldsymbol{c})$ (resp. $x = \frac{\phi(\boldsymbol{b}) - \phi(\boldsymbol{a})}{\phi(\boldsymbol{d}) - \phi(\boldsymbol{c})}$).

It is important to note that, while we used an additive representation to state the **Pluralistic Tradeoffs Account**, this was an arbitrary choice (Krantz et al. 1971; Weymark 2005; Nebel 2024). The account in no way entails that additive representations are somehow privileged. Given an alternative non-additive representation ψ, it will not be true that the well-being difference between a and b is at least as great as that between c and d iff $\psi(\boldsymbol{b}) - \psi(\boldsymbol{a}) \geq \psi(\boldsymbol{d}) - \psi(\boldsymbol{c})$. For example, if $\psi = e^{\phi}$, then we would need to state the account as follows: the well-being difference between lives a and b is at least as great as (resp. x-times as great as) that between c and d iff $\frac{\psi(\boldsymbol{b})}{\psi(\boldsymbol{a})} \geq \frac{\psi(\boldsymbol{d})}{\psi(\boldsymbol{c})}$ (resp. $x = \ln \frac{\psi(\boldsymbol{b})}{\psi(\boldsymbol{a})} : \ln \frac{\psi(\boldsymbol{d})}{\psi(\boldsymbol{c})}$).

Chapter 2

Welfare and Autonomy Under Risk

Abstract

This paper studies the relationship between promoting people's welfare and respecting their autonomy of choice under risk. I highlight a conflict between these two aims. Given compelling assumptions, welfarists end up disregarding people's unanimous preference, even when everyone involved is entirely rational and only concerned with maximizing their own welfare. Non-welfarist theories of social choice are then considered. They are shown to face difficulties, too: either they fail to respect the value of welfare in at least one important sense, or they end up prioritizing different people's welfare differently in non-risky choices, on the basis of their attitudes to risk, which are intuitively irrelevant in this context.

2.1 Introduction

Everyone agrees that the well-being of individuals matters. Welfarists believe it is all that matters. Welfarism is commonly assumed in social choice theory, and has many proponents in ethics.[1] The aim of this paper is to study the connection between the value of welfare and that of respecting people's autonomy of choice under risk. I present a framework for carrying out this task, and then offer two impossibility theorems about the connection between welfare and autonomy. First, I will show that, given compelling assumptions, any welfarist theory is committed to disrespecting the unanimous preference of individuals for a given policy over another. This will pave the way for an exploration of non-welfarist theories of social choice that respect people's autonomy in conditions

[1]For the significance of welfarism in social choice theory, see, e.g., Mongin and d'Aspremont (1998, p. 394) and Adler (2019, p. 27); for discussions of welfarism in ethics, see, e.g., Moore and Crisp (1996) and Keller (2009).

of risk. For illustration purposes, I will consider two classes of non-welfarist theories, and discuss distinct problems that they each face; I will then show that there is no non-welfarist theory that avoids both of those types of problem.

Let me introduce in more detail my contribution by situating it within two broad traditions. First, there is a variety of well-established objections to welfarism based on appeals to the value of autonomy.[2] Different authors are interested in different facets of this value, but *non-interference* is usually a common focus. Famously, Sen (1970), (1979) highlights a tension between welfarism and a certain liberal value of individual discretion over personal choices. Sen shows that a unique concern for welfare in social choice rules out the possibility of a sphere of private concern over which people have freedom to act as they wish. Nozick (1974) pursues a distinct but related line of thought when he argues against "patterned" principles of distributive justice—which include all welfarist theories—on the basis that no such principle "can be continuously realized without continuous interference with people's lives" (Nozick 1974, p. 163), since "liberty upsets patterns" (Nozick 1974, p. 160).

Here, I will be interested in an aspect of the value of autonomy that is only loosely connected to non-interference, and refers instead to a form of deference to people's own attitudes towards risk. This notion comes from the second broad tradition that I shall address, which is the literature on social choice under risk.[3] The main question in this area is as follows: suppose we need to choose one of various possible policies, each of which will affect the welfare of a number of individuals. We are uncertain about the outcome that each policy will bring about, but can assign probabilities to the different possible states of the world. What principles should guide our decision in such a situation?

[2]There are then authors who deny welfarism for other reasons; I shall mention two strands of literature. One rejects welfarism—at least if understood as a unique concern with "final well-being" as a measure of "how well [a person's] life goes" (Voorhoeve and Fleurbaey 2016, p. 936)—because of concerns with the *fairness* of how a certain distribution of welfare is attained—see, e.g., Voorhoeve and Fleurbaey (2016) and Voorhoeve (2021). The other rejects welfarism to take account of equality of opportunity, where social evaluation is based on information relating to an individual's *effort* as well as on how well off they are—see Ferreira and Peragine (2016) for an overview.

[3]For overviews of this literature, see Mongin and Pivato (2016), Fleurbaey (2018), and Adler (2019).

One of the most widely discussed principles encapsulates the ideal of autonomy that will be our focus: it roughly says that, if everyone involved prefers a certain policy over another, then we shouldn't choose the latter. Versions of this thought are usually referred to as the *ex ante Pareto* principle. I shall later qualify this principle, and make it precise. What is important to note now is that this principle, together with *dominance* reasoning, gives rise to a range of striking impossibility results. Dominance reasoning connects the evaluation of acts to that of outcomes—the principle of statewise dominance, in particular, says that if an act results in at least as good an outcome as another act under every possible state of the world, then the former act is at least as good as the latter. The impossibility results of interest here all trace back to Harsanyi 's (1955) aggregation theorem, and build upon it in various ways.[4] A strengthened version of Harsanyi's result shows the following (Fleurbaey 2009, 2010). Assume that individual preferences over policies satisfy the axioms of the orthodox view in decision theory, expected utility theory. Then, under weak assumptions, ex ante Pareto and statewise dominance entail that policies must be ranked at the social level by a sum of individual decision-theoretic "utilities"—these are the numbers representing the quantity whose expectation individuals maximize.[5] When this quantity is equated with welfare, Harsanyi's aggregation theorem becomes an argument for utilitarianism, and this is indeed how Harsanyi himself interpreted his theorem (Harsanyi 1977, 1982). The equation of decision-theoretic utility and welfare means that individuals are expected welfare maximizers, an assumption that has come to be known as Bernoulli's Hypothesis, or *Bernoulli* for short (Broome 1991b; Adler 2019).

Some recent work shows that if we reject Bernoulli by relaxing expected utility theory and allow individuals to have heterogeneous preferences for policies, the conflict between ex ante Pareto and statewise dominance is not solved, but rather exacerbated: those who evaluate outcomes in a

[4] A list that is not exhaustive but representative for our purposes includes Harsanyi (1977), (1982), Broome (1991b), Weymark (1991), Mongin (1994), Fleurbaey (2009), (2010), Blackorby, Donaldson, and Weymark (2008), Mongin and Pivato (2015), Danan, Gajdos, and Tallon (2015), Blessenohl (2020), Nebel (2020), Nissan-Rozen (2021), Bradley (2022), Fleurbaey and Zuber (2022), Kowalczyk (2023), Doody (ms).

[5] To get the mathematical expectation of an uncertain quantity, we multiply the value the quantity takes in each possible state of the world by the probability of that state, and then sum all such probability-weighted values.

utilitarian fashion face it too (Blessenohl 2020; Nissan-Rozen 2021; Bradley 2022).[6] It is tempting to interpret these recent impossibilities as arguments in favor of expected utility theory. However, restoring expected utility theory is not enough. For we can relax Bernoulli while retaining expected utility at the individual level. This was in fact part of Sen (1976), (1977) and Weymark's (1991), (2005) critique to Harsanyi's interpretation of his theorem (see also Roemer 2008, pp. 141-45). On the Sen-Weymark line of thought, utilitarianism relies on a quantitative notion of welfare that is independent of an individual's preferences in conditions of risk (Sen 1986, p. 1123). Given such independently given quantities of welfare, requiring Bernoulli amounts to an unmotivated substantive assumption of risk neutrality.

This paper will build upon the Sen-Weymark observation, and offer an alternative interpretation of the impossibilities stemming from Harsanyi's aggregation theorem, which is focused on questioning welfarism.[7] In Section 2.2, I shall relax Bernoulli and leave open whether or not individuals are expected utility maximizers. I then show that, given some plausible principles involving dominance reasoning and impartiality at the level of social evaluation, ex ante Pareto entails that welfarism must be false. I will lay out a framework for social choice under risk in Section 2.3, and use it to state the conflict between autonomy and welfarism precisely in terms of a first impossibility theorem. An intuitive illustration of this result is given in Section 2.4. My interest in presenting this argument is partly to utilize it as a springboard for considering non-welfarist views that respect the value of autonomy. Section 2.5 begins an exploration of what such views might look like; I describe two different types of non-welfarist views: one ends up disregarding the value

[6]There are two other ways of relaxing Bernoulli in the context of Harsanyi-style results: one is to relax expected utility but assume that there is a homogenous ranking of policies among individuals (McCarthy, Mikkola, and Thomas 2020), the other involves relaxing expected utility theory as an assumption but deriving it back as a conclusion (Mongin and Pivato 2015; McCarthy, Mikkola, and Thomas 2020).

[7]The availability of such an interpretation is prefigured in results by Blackorby, Bossert, and Donaldson (2002, p. 587), (2005, p. 230), and Blackorby, Donaldson, and Weymark (2008, p. 161), discussed further in footnote 28; they do not, however, advocate it. Mogensen (2024) interprets the results by Blessenohl (2020), Nebel (2020), and Bradley (2022) as showing that the motivation for respecting people's preferences *ex ante* (in the context of Buchak's (2013) non-orthodox decision theory) is incompatible with consequentialism—cf. Frick (2013, p. 132). In this paper, I retain instead a broadly consequentialist framework and pin the blame on the claim that only welfare matters to the social evaluation of outcomes. Blessenohl (2020, pp. 505-506) briefly mentions a non-welfarist response to his impossibility result, a version of which I took up and defended in Cibinel (2022). My aim here is to explore this suggestion at a much greater level of generality.

of welfare altogether, while the other gives priority to people's welfare on the basis of their risk attitude. Section 2.6 generalizes the point via a second impossibility theorem, showing that any autonomy-based non-welfarist theory entails some version of one of these two implications. Section 2.7 concludes by considering different reactions one might have to the results of the paper.

Before we delve into the details, it will be useful to have a rough version of the conflict between welfarism and autonomy on the table; I illustrate it now with an example. There are only two individuals affected, and we need to choose between one of two policies. Policy 1 gives 10 units of welfare to the first individual, and 0 to the second, if event E occurs.[8] It gives 0 units of welfare to the first individual, and 10 to the second, if event $\neg E$ occurs. Each of E and $\neg E$ have probability 0.5 of occurring. Policy 2 gives 5 units of welfare to both individuals, no matter whether E or $\neg E$ obtains. Perhaps, both individuals are suffering from a moderate illness (welfare level 0), and there is a medicine that will fully cure the first individual (welfare level 10) and leave the second ill if E obtains, and fully cure the second individual and leave the first ill if $\neg E$ obtains. However, a second medicine will guarantee partial recovery for both (welfare level 5). We can represent the case in matrix form, as follows.[9]

	E	$\neg E$
policy 1	(10, 0)	(0, 10)
policy 2	(5, 5)	(5, 5)

Figure 2.1

Each cell in the matrix represents a certain final outcome that might obtain; a pair of numbers (r_1, r_2) stands in for the first individual getting r_1 units of welfare, and the second individual r_2

[8]The example makes certain assumptions about welfare measurement that are carefully considered only in Section 2.2.

[9]This kind of case is very often used to illustrate the puzzles raised by Harsanyi's aggregation theorem; see, e.g., Fleurbaey and Voorhoeve (2013, p. 116), Frick (2015, p. 191), Nissan-Rozen (2017), and Doody (ms). The connection it exemplifies between risk aversion and equality on the one hand, and risk-seekingness and inequality on the other, predates even Harsanyi's aggregation theorem—see, e.g., Friedman (1953, p. 278). See also the literature on "option luck," concerning whether individuals should be held responsible for the results of "deliberate and calculated gambles" (Dworkin, 1981, p. 293). Footnote 28 below compares my result based on this case, Theorem 1, and additional related work—including Hammond (1981), Blackorby, Donaldson, and Weymark (2008), and Fleurbaey and Zuber (2021b), (2022).

units of welfare, in the given outcome. Welfarists think that well-being is all that matters to evaluating outcomes. They say: no matter what other circumstances obtain, there is a fixed fact of the matter as to whether, for example, the outcome of policy 1 under E is better than the outcome of policy 2 under E. The matter is fixed, according to welfarists, because we know how well off each individual would be in those outcomes. Since nothing else matters, this is enough to evaluate them.

However, on pain of violating people's autonomy, we can show there *are* additional circumstances that should affect our evaluation of outcomes. One way to fill in the details of the case might be this. Suppose both individuals are risk-seeking and prefer policy 1 to policy 2. By the ex ante Pareto principle, policy 1 is indeed better than policy 2 on this version of the case. But, plausibly, policy 1 can't be better unless there is some state in which its outcome is better than that of policy 2. Moreover, clearly the outcome of policy 1 under E is exactly as good as the outcome of policy 1 under $\neg E$. So, on this version of the case, both of the possible outcomes of policy 1 are better than the single outcome that policy 2 is sure to result in.

Here is a different way to fill in the details of the case. Suppose both individuals are risk-averse and prefer policy 2 to policy 1. By the ex ante Pareto principle, policy 2 is indeed socially better than policy 1 on this version of the case. But, plausibly, policy 2 can't be better unless there is some state in which its outcome is better than that of policy 1. Moreover, since the two possible outcomes of policy 1 are equally good, the single outcome that policy 2 is certain to result in must be better than both, on this version of the case.

Under one set of circumstances, an outcome with distribution of welfare $(10, 0)$—or $(0, 10)$—is better than one with distribution of welfare $(5, 5)$. Under another set of circumstances, the outcome with welfare distribution $(5, 5)$ is better than that with welfare distribution $(10, 0)$—or $(0, 10)$. The facts about welfare remain the same, but our evaluation changes. Therefore, welfarism is false. Very informally, this is the argument that motivates the study that follows.

2.2 Welfare and Risk

This section clarifies the assumptions about welfare and risk attitudes that the argument just presented relies upon. We begin with a qualification on ex ante Pareto. As I said, this principle gives voice to a certain ideal of autonomy, to do with respecting people's attitudes to risk when their welfare is at stake. My earlier, first pass formulation was this:

> if every individual involved prefers a certain act or policy over another, then we shouldn't choose the latter.

Without qualification, this principle is dubious. Suppose the individuals involved prefer a certain act over another, but we know that this is because they have false beliefs, either about their environment or about what their good consists in. Even if we don't know that their beliefs are false, we might know that they are based on much less evidence than we possess, and that their preferences would change if they knew more. Or perhaps we all share the same beliefs and evidence, but these individuals' preferences are in some sense irrational. A staunch opponent of paternalism might claim that even so, we should defer to what these people want. But some would concede that in such circumstances people's preferences may legitimately be overridden.[10] I will leave this issue to one side. I am concerned with a special kind of case, in which, I think, failing to respect people's preferences is much more problematic.

Suppose every individual affected shares the same evidence, and they all share the same evidence with us, who are engaged in the exercise of ranking policies. All of our beliefs are modeled by a single, given probability distribution over the possible states of the world. Moreover, the individuals affected know exactly what would be ultimately good for them, always and only preferring outcomes in which they get more welfare (we needn't make the rather implausible assumption that everyone is self-interested, but only the more innocuous one that self-interested preferences are the legitimate input to social choice). Individuals' preferences over policies with uncertain conse-

[10] See Broome (1987, p. 408) for an early statement of the problems that arise for ex ante Pareto when probabilities differ; cf. Mongin (2016), who suggests that ex ante Pareto should be rejected in all cases of "spurious unanimity," in which (roughly) people's preferences all agree but are based on different reasons.

quences are rational, in the most stringent sense of the term. First, they are coherent in the sense of satisfying the axioms of the correct decision theory. Second, they are reasonable in the sense of not being in any other way criticizable from the perspective of someone who is uniquely concerned with their interest.[11] In this context at least, the ex ante Pareto principle is very hard to resist. It is this restricted version of the principle that I shall use throughout the paper.

I shall assume that we can compare the welfare levels and improvements of different people on a single scale, before any apparatus for dealing with risk is introduced—this assumption corresponds to taking seriously the Sen-Weymark challenge to Harsanyi mentioned in Section 2.1. In the technical jargon, welfare is cardinally measurable and fully comparable. What this means is that it makes sense to say things like "person 1's welfare in some outcome is greater than person 2's welfare in another" and "person 1 would benefit more if we acted in this way than person 2 would if we acted in that way." These assumptions make it possible to represent welfare levels with numbers.[12] When, below, I say that a certain individual gets welfare level x in a certain outcome, with x a real number, what I mean is that this individual gets x units of welfare as measured by a scale appropriately chosen to reflect the assumptions about welfare that I just described (i.e., what is called an *interval scale*, common among all individuals).

These assumptions also entail that it makes sense to speak of risk neutrality, risk aversion, and risk-seekingness in relation to welfare. I will call someone *risk-neutral* if and only if they are always indifferent between policies that yield the same expectation of welfare for them, *risk-averse* if and only if they always prefer a policy that gives them x units of welfare for sure to a policy with uncertain consequences that gives them x units of welfare in expectation, *risk-seeking* if and only if they always prefer a policy with uncertain consequences that gives them x units of welfare in

[11] For this second sense of rationality, I draw on Parfit (2011, p. 33). Buchak (2017a, p. 620) and Hájek (2021, p. 190) specifically give a version of this distinction in relation to attitudes to risk. Those skeptical of there being anything more to rationality than coherence can simply disregard this point: my assumption, for them, is just that all individuals are rational in the sense of having preferences that are coherent.

[12] See Krantz et al. (1971, ch. 4), for the relevant representation theorems; Adler (2019, p. 268) and Nebel (2024) specifically apply Krantz et al.'s results to the issue of welfare measurement.

expectation to a policy that gives them x units of welfare for sure.[13] Orthodox decision theory and its rivals alike make room for risk neutrality, risk aversion, and risk-seekingness for welfare (see Broome 1991a; Buchak 2013; Bottomley and Williamson 2024). So there is no doubt that both risk-averse and risk-seeking preferences, as well as risk-neutral ones, can be rational in the sense of satisfying the axioms of the correct decision theory. All I will assume about the structure of a person's preference relation over policies is that it is an ordering (namely, it is reflexive, complete, and transitive); this makes what I have to say compatible with orthodox decision theory, but also with any of its main rivals.

A further question one might ask is which patterns of preferences for risk are reasonable, in the sense of being immune from any kind of criticism from the perspective of someone concerned with the person's interest. We have already assumed that people always prefer more welfare to less. The question is whether there is any prudential advantage that one particular attitude—for example, risk neutrality—possesses over the others. I shall assume that the answer is "no." Perhaps extreme levels of proneness to risk, as well as extreme levels of risk aversion, can be ruled out on this basis. But there is certainly a range of entirely reasonable preferences for risk involving one's welfare.[14] I will call this thesis *welfare-risk permissivism*.

The thesis of welfare-risk permissivism gives a positive spin to the negative observation of Sen and Weymark, according to which Harsanyi is unwarranted in assuming Bernoulli. Before moving on, we should briefly consider two related attempts to salvage Bernoulli. The first strategy claims that we do not in fact possess any independently given quantitative notion of welfare, relative to which individuals can display differing risk attitudes. The suggestion would then be that equating decision-theoretic utility with welfare allows us to give meaning to quantitative statements

[13] This is a standard characterization of risk preferences with respect to a given quantity—see, e.g., Buchak (2013, p. 21-22) for a discussion of it and of related characterizations. Note also that my definitions of risk aversion and risk-seekingness are "global"; it is worth noting that my results below actually only rely on risk aversion and risk-seekingness being displayed "locally" with respect to some interval of welfare levels.

[14] Versions of this thought are articulated by Buchak (2017a), Hájek (2021), and (on one reading) Broome (1991a, p. 6). Broome (1991b), however, ends up accepting a version of Bernoulli, which requires reasonable people to be expected welfare maximizers. I will discuss this issue presently.

concerning welfare, which would otherwise make no sense or display semantic indeterminacy (Broome 1991b, 2004; Risse 2002; Greaves 2017). The second strategy claims that, while we do have some quantitative notion of individual "benefit" prior to choice under risk, an individual's *welfare* is nonetheless affected not only by what benefit they receive, but also by the probability that was associated with receiving that benefit.[15] An individual may be risk-averse with respect to what benefits they get, but should come out as risk-neutral with respect to welfare itself.

Both of these strategies, however, deliver counterintuitive verdicts, owing to the fact that risk attitudes seem irrelevant to determining how well off an individual is (see Fleurbaey and Zuber 2022). To illustrate, suppose you are uniquely concerned with a particular individual's welfare. You don't know whether this individual is risk-averse or risk-seeking with respect to experiencing pain but you do know that they ended up in some state of pain. Would you rather learn that they are risk-seeking and in more pain, or risk-averse and in less pain? On both of the strategies for salvaging Bernoulli just considered, there will be attitudes to risk and states of pain such that you should be indifferent between which of the two pieces of news you'd rather learn.[16] But if all you genuinely care about is this person's welfare, you should be relieved to know they are in less pain, and indifferent about what risk attitude they have. So, you should want the individual risk-averse and in less pain. To the extent that we might be inclined to prefer them risk-seeking and in "more pain," it seems to me that this is because it comes natural to associate risk-seekingness with a greater ability to endure misfortune. But this is not how we are supposed to think about this example, for surely such an ability would affect how much pain one feels. Notice, moreover, that while the example is specifically about pain, the point it makes does not rely on assuming hedonism about welfare.

[15]This strategy could be developed by appealing to Stefánsson and Bradley's (2015) denial of what they call "chance neutrality"; see also Goldschmidt and Nissan-Rozen (2021) and Bradley and Stefánsson (2023). Cohen et al. (2022, p. 2562) specifically show how such a strategy can help respond to Nissan-Rozen's (2021) impossibility, which I mentioned above. I thank a referee for inviting me to consider this view.

[16]On the first strategy, since the shape of the individual's expected utility function is different depending on whether they are risk-averse or risk-seeking with respect to pain, equal increases in pain can lead to different welfare losses. See Adler (2019, pp. 55-64) and Fleurbaey and Zuber (2021a) for more on this topic. Below we will consider a view where expected utility is not a measure of welfare differences, but rather of the moral significance of welfare differences—see Section 2.5.1 and especially footnote 29 (compare Nebel 2022a, p. 24). On the second strategy, if we assume that the pain resulted from a risky option, the welfare loss that the risk-seeking individual gets from the risk (i.e. the antecedent probability of that pain) will be lower than for the risk-averse (and may even be a gain, see Stefánsson and Bradley 2015, p. 605). This will make up for the added pain the risk-seeking individual experiences.

We could instead suppose that the individual fares better, if risk-averse, along any other objective good (such as friendship, knowledge, and so on) or with respect to desire satisfaction and strength of preference (provided that the latter is measured independently of risk, which can be done along the lines indicated by Bell and Raiffa 1988).

This argument no doubt falls short of a conclusive refutation of the strategies considered for salvaging Bernoulli.[17] But it does at least suggest that following the Sen-Weymark critique and severing our notion of quantities of welfare from that of risk attitudes has intuitive appeal. The study of the paper can then be seen as exploring the consequences of this approach for those who want to both promote welfare and respect autonomy under risk.

2.3 Framework and First Result

I now lay out a framework that will allow us to state the conflict between autonomy and welfarism more precisely. The framework I use combines a simplified version of Sen's (2017) approach of *social welfare functionals* with a minimal apparatus for choice under risk.[18] To reiterate, the key component of our approach will be that cardinal facts about welfare are given; people's preferences between policies are then formed on the basis of what quantities of welfare they would receive from each policy, with which probabilities.

We assume that there is a set N of n individuals whose welfare our choices might affect, with $n \geq 2$. There is a set X of outcomes, each of which includes all information relevant for purposes of ethical evaluation: any outcome will certainly tell us about every individual's welfare, but we

[17] **Chapter 1** of this dissertation, of course, considered in much greater detail how quantities of welfare can be obtained independently of risk; see Dietrich (ms) for a recent proposal on how to disentangle welfare from decision-theoretic utility.

[18] Two early frameworks that make use of Sen's approach in the context of choice under risk are given by Mongin (1994) and Mongin and d'Aspremont (1998). For approaches more similar to mine, see Blackorby, Bossert, and Donaldson (2005, ch. 7), Blackorby, Donaldson, and Weymark (2008), Chambers and Echenique (2012) and Fleurbaey and Zuber (2021a)—some differences are highlighted in footnotes below. Another connection is with frameworks that explicitly make room for a non-welfare component, if only to better characterize welfarism (see Blackorby, Bossert, and Donaldson 2005, ch. 3, 2006). The non-welfare component, in my framework, will be given by the apparatus for choice under risk.

leave open that it may contain other information, too, since we don't want to prejudge the question of welfarism. Information specifically about welfare is captured by a fixed welfare function W, which associates to each outcome a list of welfare levels: the first entry in the list is a number representing the welfare level the first individual gets in that outcome, the second entry is a number representing the welfare level of the second individual, and so on.[19] So, for example, with two individuals, there may be some outcome x such that $W(x) = (10, 0)$. To single out the i^{th} entry in this outcome's list of welfare levels, we write $W_i(x)$. Individual preferences over outcomes in X are encoded in the welfare function W, given our assumption that all individuals ultimately care about is improving their welfare—or, at least, that such self-interested concern is the legitimate input to social choice.

To model uncertainty, let S be a finite set of k states of nature, and let $\pi(s)$ be the probability of state s. This fixed probability function reflects the assumption I made that everyone shares the same evidence and beliefs. I will assume that S can be partitioned into two equiprobable events, E and $\neg E$; these could be, for instance, the two possible results of a fair coin flip. There is a set A of policies or acts, which we model as functions from states of nature to outcomes. Crucially, we do *not* make the following common assumption (following Broome's nomenclature, 1991b, p. 80).

RECTANGULAR FIELD: A is the set of all functions from states to outcomes.

In other words, we leave open that only some of the functions from states to outcomes stand in for genuine policies we may implement. The reason for this will emerge in due course, but is roughly the following: it is important for some of the theories I will consider below that outcomes include information about how they come about. If an outcome x results from a genuinely risky act, and such information is included in the outcome itself, it would indeed make little sense for there to be a function in A that returned x in all states.

[19] Our assumption of a fixed welfare assignment should be acceptable for those who think that there couldn't be a welfare difference for an individual without a difference in the outcome that individual gets. Those who reject this claim should see the choice of a fixed welfare profile as a simplification to limit the number of variables in my framework; all my results could be restated in a more general setting with multiple welfare profiles (following Sen (2017) more closely).

Each individual i associates to each act a a prospect a_i, which tells us how much welfare i gets in each of the states, if act a is chosen.[20] Individuals have preferences over prospects, encoding their risk attitudes: for each individual i, there is a reflexive, complete, and transitive relation R_i—that is, a "weak preference" ordering—over the prospects A offers to i. Strict preference and indifference are understood in the usual way and denoted by P_i and I_i respectively.[21] I assume that individual preferences are defined over the same set of prospects, so that it makes sense to say that two individuals share the same preference relation or risk attitude.[22] Notice that, when a prospect results in a given outcome with certainty, the individual's ranking of that prospect is determined by how much welfare it gives to that individual: the more, the better.

These orderings over prospects R_i are grouped together by an n-tuple $R = (R_1, ..., R_n)$, which we call a risk profile, or simply a profile. The set of all rational risk profiles is denoted by $\mathcal{R}$. A risk profile is rational if it is made up entirely of rational orderings, both in the sense of rationality as coherence and in the sense of rationality as reasonableness. We can now state formally the claim that there are multiple rationally permissible attitudes to risk involving one's welfare.

> WELFARE-RISK PERMISSIVISM: $\mathcal{R}$ includes profiles with risk-neutral, risk-averse, and risk-seeking orderings of prospects.

Notice that all results below actually only need there to be two distinct orderings—for example, one risk-averse and one risk-seeking. With this setup in place, we can formulate the problem of social choice under risk as follows: given any profile R, we want to be able to rank acts and outcomes. Our primary focus will be on the following kind of functional relations:

- A *social quasi-ordering function* is a function that maps each profile $R \in \mathcal{R}$ to a pair of reflexive and transitive social preference relations, i.e. quasi-orderings, $(\succeq_R^A, \succeq_R^X)$, with $\succeq_R^A$ over A and $\succeq_R^X$ over X.

[20] Formally, for all states $s \in S$, $a_i(s) = W_i(a(s))$.

[21] For all acts $a, b \in A$, if $a_i R_i b_i$ and not $b_i R_i a_i$, we write $a_i P_i b_i$. If both $a_i R_i b_i$ and $b_i R_i a_i$, we write $a_i I_i b_i$.

[22] This imposes some constraints on the sets X of outcomes and A of acts. In particular, if $i, j \in N$ and $a \in A$ are fixed, there must be $b \in A$ such that $a_i = b_j$. In turn, this has implications for what outcomes exist. However, this assumption is uncontroversial if we think of acts as *possible* policies a decision-maker might face, and of outcomes as merely possible final states of affairs.

Strict preference and indifference over acts are denoted by $\succ_R^A$ and $\sim_R^A$, each again defined in the standard way; the analogous convention holds for preference over outcomes.[23]

A note about the need for two social preference relations, one over acts and another over outcomes. Had I assumed Rectangular Field above, I could have defined the social preference relation over outcomes in terms of how $\succeq_R^A$ ranks *sure acts*, that is, acts that return the same outcome in every state of the world. But I did not make this assumption, so there is no guarantee that, given some outcome x, we will find an act a in A that returns x in every state. If risk has a certain kind of normative relevance, there might be good reasons to think that such a search should systematically fail, as briefly hinted at above and explained more carefully below (see also Broome 1991b, pp. 115-117). For readability, however, I will omit the superscripts when no confusion arises.

We can finally address the problem of social choice under risk: how should we rank acts and outcomes? We impose three axioms on social quasi-ordering functions. For each axiom, I will first describe it informally, and then give a formal statement. The first axiom is a package of dominance principles, which guarantees that our evaluation of acts coheres with our evaluation of outcomes. This axiom has two parts. Part (i) says: if an act is better than another socially speaking, then the outcome of the former is better than the outcome of the latter in at least some possible state of the world. And part (ii) says: if the outcome of an act is socially better than the outcome of another act in every state, then the former act is socially better than the latter.

> DOMINANCE PACKAGE: For all profiles $R \in \mathcal{R}$, and for all acts $a, b \in A$, (i) if $a \succ_R b$ then there exists $s \in S$ such that $a(s) \succ_R b(s)$; (ii) if $a(s) \succ_R b(s)$ for all $s \in S$, then $a \succ_R b$.[24]

[23] For all $a, b \in A$, and $R \in \mathcal{R}$, if $a \succeq_R^A b$ and not $b \succeq_R^A a$, we write $a \succ_R^A b$; if both $a \succeq_R^A b$ and $b \succeq_R^A a$, we write $a \sim_R^A b$. The analogous convention holds for $\succeq_R^X$.

[24] Some (e.g., Hare 2010; Bader 2018) will be suspicious of part (i) of Dominance Package because, given incompleteness, it conflicts with stochastic dominance in cases of "opaque sweetening," for which see Hare (2010). While I do not assume completeness for social preference, this needn't worry us. Variants of my results could be proved using the following weaker principle, which Hare's case gives us no reason to doubt: if a is better than b then *either* a stochastically dominates b *or* there exists a state in which the outcome of a is better than that of b. For my first result, alternatively, Lederman's (forthcoming) "negative dominance" could be employed instead. I use Dominance Package (i) to avoid complicating my discussion and proofs.

Next, we have the ex ante Pareto principle discussed in Section 2.2. This principle ensures that if everyone affected prefers their prospect with some act when compared to their prospect with another act, then the former act is indeed socially better than the latter. Formally, we have

> EX ANTE PARETO: For all profiles $R \in \mathcal{R}$ and acts $a, b \in A$ such that $a_i \neq b_i$ for some $i \in N$, if $a_i P_i b_i$ for all i such that $a_i \neq b_i$, then $a \succ_R b$.

The third principle is required for impartiality, or equal concern for all individuals. Say that two acts a and b are permutations of one another if they distribute the very same prospects, though possibly to different people.[25] In words, the principle states that when everyone shares the same preferences for risk, and two acts are permutations of one another, then it can't be that one of the two acts is better than the other.

> HOMOGENOUS ANONYMITY: For all profiles $R \in \mathcal{R}$ and acts $a, b \in A$, if a and b are permutations of one another, and $R_i = R_j$ for all $i, j \in N$, then neither $a \succ_R b$ nor $b \succ_R a$.

Our three axioms constrain social quasi-ordering functions in different ways. Dominance Package constrains social quasi-ordering functions with respect to how they relate social preference over acts in A to social preference over outcomes in X, in any given profile. Ex Ante Pareto and Homogenous Anonymity constrain social quasi-ordering functions with respect to how they rank acts in any given profile. When combined with Dominance Package, these two axioms are capable of entailing verdicts about social preferences over outcomes. Notice, however, that we do not assume any axiom directly about social preference over outcomes.

Lastly, we need to characterize welfarism. In social choice theory, the welfarist stance is regimented in terms of the thesis that "[...] the ethical ranking of outcomes is determined by individual well-being. [...] It takes a well-being difference to make an ethical difference in the outcomes" (Adler 2019, p. 27). In the current framework, this thesis can be formulated as follows: there is a single reflexive and transitive relation on lists of welfare levels that fully determines the so-

[25] Formally, there is a bijection $\sigma : N \to N$ such that for all $i \in N$, and for all $s \in S$, $W_i(a(s)) = W_{\sigma(i)}(b(s))$.

cial ranking of outcomes. Formally, call $\mathcal{W}$ the subset of $\mathbb{R}^n$ (which is the set of all n-tuples of real numbers) such that $(r_1, ..., r_n) \in \mathcal{W}$ just in case there exists an outcome x in X with $W(x) = (r_1, ..., r_n)$. Then we have

> WELFARISM: There exists a unique quasi-ordering $\succeq$ on $\mathcal{W}$ such that for all outcomes $x, y \in X$ and for all profiles $R \in \mathcal{R}$, $x \succeq_R y$ if and only if $W(x) \succeq W(y)$.[26]

We can better understand the commitments of Welfarism by noticing that it entails two distinct principles. The first, Pareto Indifference, says that—keeping the risk profile fixed—two outcomes associated with the same list of welfare levels must be ranked as equally good. The second, Independence of Irrelevant Alternatives, says that changing non-welfare aspects of a situation should not affect the ranking of outcomes. In our framework, this means that tweaking what risk attitudes individuals have should not make a difference to how outcomes are assessed: a social quasi-ordering function satisfying Independence of Irrelevant Alternatives assigns the *same* quasi-ordering over outcomes to every profile in $\mathcal{R}$. Formally, we have (cf. Bossert and Weymark 2004, pp. 1105-1106):

> PARETO INDIFFERENCE: For all outcomes $x, y \in X$ and for all profiles $R \in \mathcal{R}$, if $W(x) = W(y)$, then $x \sim_R y$.

> INDEPENDENCE OF IRRELEVANT ALTERNATIVES: For all outcomes $x, y \in X$ and for all profiles $R, R' \in \mathcal{R}$, $x \succeq_R y$ if and only if $x \succeq_{R'} y$.[27]

Since these two principles are entailed by Welfarism, if either of them fails Welfarism does, too.

[26] Cf. Bossert and Weymark (2004, p. 1107). Standard formulations often assume that the "welfarist" relation $\succeq$ is complete; since every ordering is a quasi-ordering, but not the other way around, the formulation I use here is more general.

[27] Chambers and Echenique (2012, p. 587) and Fleurbaey and Zuber (2017, p. 680) consider a principle structurally identical to Independence of Irrelevant Alternatives, which they call "Invariance to risk attitudes for constant acts." Substantially, however, the two principles are different: both Chambers and Echenique and Fleurbaey and Zuber only assume, as a matter of framework, ordinal non-comparability for individual welfare. As a result, their principle is as much a restriction on well-being comparisons and measurement as it is on social evaluation. Importantly for our purposes, a denial of their principle is compatible with Welfarism. For example, they understand accepting Bernoulli as a rejection of their principle (ibid.). In my framework, where cardinal measurability and full comparability of welfare are built-in, accepting Bernoulli should be seen as claiming that $\mathcal{R}$ only includes the risk-neutral ordering of prospects. Bernoulli thus entails Independence of Irrelevant Alternatives.

We are now in a position to present the first main result of the paper formally: in our framework, Welfarism is incompatible with the conjunction of Dominance Package, Ex Ante Pareto, and Homogenous Anonymity—given a technical assumption about the sets of outcomes and policies which I will call Domain Condition.

> **Theorem 1.** *Given Welfare-Risk Permissivism and Domain Condition, no social quasi-ordering function satisfies Dominance Package, Ex Ante Pareto, Homogenous Anonymity, Pareto Indifference, and Independence of Irrelevant Alternatives.*[28]

The proof of Theorem 1 is given in the Appendix, together with a statement of the Domain Condition, which is of little philosophical import. Essentially, by combining Ex Ante Pareto and Homogenous Anonymity (which concern social preference over acts) with Dominance Package (which connects social preference over acts with social preference over outcomes) we get verdicts about social preference over outcomes that contradict Welfarism (which concerns social preference over outcomes). The next section explains the impossibility that Theorem 1 gives rise to, giving a more detailed exposition of the argument sketched in Section 2.1.

[28]Here is a comparison with relevant extant literature. First, results pitting welfarism against ex ante Pareto principles given a rejection of Bernoulli in a multi-profile framework appear in Blackorby, Bossert, and Donaldson (2002, p. 587), (2005, p. 230), and Blackorby, Donaldson, and Weymark (2008, p. 161). Besides differences in framework, these results assume a version of rectangular field and expected utility theory for individual and social preference, while mine does not; I allow, moreover, for incomplete social preferences. Theorem 1 relies instead on the weaker principle of Dominance Package and on Homogenous Anonymity, with a minimal domain condition. Second, the conflict my result points to is reminiscent of the debate spawned by Harsanyi on whether the "social welfare function" should be applied *ex ante* or *ex post*; see especially Hammond (1981), (1982). The main difference here is that my result is multi-profile: in this context, welfarism corresponds to the idea that the social welfare function *ex post* should not depend on what profile we're in. Denying welfarism in this sense is compatible with a range of approaches to the *ex ante-ex post* debate, including a rapprochement of the two perspectives, with the ex post social welfare function differing in each profile (Section 2.5.1), and a denial of the ex post approach, which however retains dominance reasoning (Section 2.5.2). Finally, my theorem can be related to results by Fleurbaey and Zuber (2021b), (2022): they show that social evaluation cannot both comply with statewise dominance and a weak version of ex ante Pareto while not paying attention to risk attitudes in the absence of risk (see their Theorem 1 in Fleurbaey and Zuber 2021b, p. 14, and their Theorem in Fleurbaey and Zuber 2022, p. 459). Formally, their results rely on social evaluation *across* profiles, e.g. "outcome x with profile R is better than outcome y with profile R'." My result, by contrast, only relies on intraprofile social evaluation. The proof strategy is nonetheless similar; each proof proceeds by considering two profiles with homogenous but different risk attitudes. At the philosophical level, they do not interpret their results as questioning welfarism. This is because, in their framework, risk attitudes are *not* with respect to independently given and interpersonally comparable quantities of welfare (as they are in mine). So, they take a denial of their principles forbidding risk attitudes to count in the absence of risk to mean that risk attitudes must play a role in the measurement of individual welfare (Fleurbaey and Zuber 2021b, p. 25, 2022, p. 463). They do not consider the alternative route, which my Theorem 1 is precisely supposed to highlight, of claiming that while risk attitudes do not affect individual welfare measurement, they do affect social evaluation in a way that is inconsistent with welfarism.

2.4 A More Detailed Illustration

Suppose there are only two individuals, and that they share the same ordering of prospects. For ease of exposition, I assume completeness (i.e., that $\succeq_R^A$ and $\succeq_R^X$ are orderings); Theorem 1 shows it is not actually needed. Consider first a choice between the following two sure acts, a and b, where s^* is the disjunction of all states.

	s^*
a	(10, 0)
b	(0, 10)

Figure 2.2

Act a gives 10 units of welfare to individual 1 and 0 to individual 2 for sure, and b gives 0 units of welfare to individual 1 and 10 to individual 2 for sure. a and b are thus permutations of one another. By Homogenous Anonymity, neither is better than the other. By the contrapositive of Dominance Package (ii) and completeness, a welfarist must conclude that:

> $(*)$ any outcome with welfare distribution $(10, 0)$ is exactly as good as any outcome with welfare distribution $(0, 10)$.

Consider next a second choice, a copy of the one introduced in Section 2.1.

	E	$\neg E$
c	(10, 0)	(0, 10)
d	(5, 5)	(5, 5)

Figure 2.3

E and $\neg E$ partition S, the set of possible states, into two equiprobable events. By Welfare-Risk Permissivism, we may assume there is some profile $R \in \mathcal{R}$ such that both individuals prefer c to d, and some profile $R' \in \mathcal{R}$ such that both individuals prefer d to c. Consider first R. Because both individuals prefer c to d in R, by Ex Ante Pareto $c \succ_R d$. But then by Dominance Package (i) and $(*)$, a welfarist must accept that:

($\dagger$) any outcome with welfare distribution $(10, 0)$ is better than any outcome with welfare distribution $(5, 5)$, and any outcome with welfare distribution $(0, 10)$ is better than any outcome with welfare distribution $(5, 5)$.

However, consider next profile R'. Because both individuals prefer d to c in R', by Ex Ante Pareto we have $d \succ_{R'} c$. Then, by Dominance Package (i) and $(*)$, a welfarist must conclude that any outcome with welfare distribution $(5, 5)$ is better than any outcome with welfare distribution $(10, 0)$ and also better than any outcome with welfare distribution $(0, 10)$. But this contradicts $(\dagger)$.

Since all of Dominance Package, Ex Ante Pareto, and Homogenous Anonymity are very plausible, the above argument can be taken as an objection to Welfarism. The particular aspect of Welfarism to be blamed is Independence of Irrelevant Alternatives. For we can avoid contradiction by accepting that in some cases (under profile R, where both individuals prefer c to d) the outcome with welfare distribution $(10, 0)$ or $(0, 10)$ is better than that with welfare distribution $(5, 5)$, but in other cases (under profile R', where both individuals prefer d to c), the outcome with welfare distribution $(5, 5)$ is better than that with welfare distribution $(10, 0)$ or $(0, 10)$. But this means that changing some non-welfare aspects of a situation—in particular, what risk attitudes people have—can make a difference to the ranking of outcomes, which is what Independence of Irrelevant Alternatives denies.

2.5 Non-Welfarist Possibilities

What does a non-welfarist theory that accommodates all of Dominance Package, Ex Ante Pareto, and Homogenous Anonymity look like? This section considers two classes of non-welfarist theories that satisfy these principles. One of these lets individual risk attitudes influence how much weight each person's welfare receives in the social ranking of outcomes. The other violates Pareto Indifference, which I see as the more innocuous implication of Welfarism. In Section 2.6, I show that this is no coincidence: in our framework, any theory that satisfies Dominance Package and Ex Ante Pareto entails one of these two *prima facie* unpalatable conclusions.

2.5.1 The Risk-Priority View

The first class of theories prioritizes individuals differently depending on their attitude to risk; I call theories in this class Risk-Prioritarian, and the general class of theories the Risk-Priority View. Risk-Prioritarian theories sometimes give priority to the welfare of more risk-averse individuals, and sometimes instead attach greater importance to the welfare of more risk-inclined people. This differential prioritization obtains because outcomes are ranked in terms of a transformed sum of individual welfare levels, where the transformation of welfare is potentially different for each person: it is a function that *also* represents this person's risk attitude.[29]

Unlike the class of theories considered in the next subsection, the Risk-Priority View gives us no reason to reject Rectangular Field, as defined above (see Broome, 1991b, p. 115). To streamline the presentation, we can thus help ourselves to this assumption—though we need to remember that it does not hold in general, but rather conditional on the Risk-Priority View. So, for each outcome there is now a "sure" act resulting in that outcome in all states.

To develop the proposal, we need more structure on the individuals' preferences over prospects. This will allow us to better measure their degree of proneness to risk. In particular, we assume that each individual i's preference relation over prospects admits of an expected utility representation U_i: a function from prospects to real numbers representing the individual's preferences, with the expected utility property—that is, the utility of an act is its expected utility. Utility functions with expectational form encode an individual's risk attitude towards their welfare (in the standard sense of Pratt 1964 and Arrow 1971) and are unique up to positive affine transformation.[30]

[29]The Risk-Priority View is the theory Harsanyi would end up with if he accommodated the Sen-Weymark critique and granted that decision-theoretic utility needn't be linear in welfare, while at the same time retaining ex ante Pareto. For precedents in the literature, see especially Broome (1991b, pp. 211-213), but also Blackorby, Donaldson, and Weymark (1980, pp. 19-31), (2008, p. 159), and Dietrich (ms, section 5). As hinted at in footnote 16, there is also an analogy here with the literature that accepts Bernoulli and addresses the question of when risk aversion in various goods should be seen as an advantage/burden with respect to one's welfare, that is, how utility/welfare should be normalized (e.g., Adler 2019, pp. 55-64; Fleurbaey and Zuber 2021a). I consider these issues in more detail in **Chapter 3** of this dissertation, where I end up defending the Risk-Priority View.

[30]So, in this sub-section I am assuming, for illustration purposes, orthodox decision theory's account of risk aversion, as presented by, e.g., Broome (1991a) and Wilkinson (2022).

Here is how the view works, in detail. Let R_* be any rational ordering of prospects an individual could have. We pick out a *privileged* expected utility representation for this ordering, and call it $\mathbf{U}_{R_*}$. Next, we do this for all other possible rational orderings of prospects. Define finally, for each profile $R \in \mathcal{R}$ and act $a \in A$,

$$U_R(a) \coloneqq \mathbf{U}_{R_1}(a_1) + ... + \mathbf{U}_{R_n}(a_n),$$

where for all individuals $i \in N$, $\mathbf{U}_{R_i}$ is the privileged expected utility representation of R_i, the ordering that i has under profile R. The Risk-Priority View then considers an act as socially better than another if and only if U_R assigns it a greater value. Notice: first, privileged expected utility representations are assigned to orderings; then, orderings are assigned to individuals given any profile $R \in \mathcal{R}$. As a result, any two individuals with the same ordering get the same privileged expected utility representation (whether within or across profiles). The ranking of outcomes immediately follows from the ranking of sure acts. If an act a results in some single outcome x in all states, we abuse notation slightly and write:

$$U_R(x) \coloneqq \mathbf{U}_{R_1}(W_1(x)) + ... + \mathbf{U}_{R_n}(W_n(x))$$

(compare Broome 1991b, p. 211). Then, outcomes with greater U_R value are better. This last equation, moreover, makes vivid the distinctive feature of the Risk-Priority View: outcomes are ranked in terms of a transformed sum of individual welfare levels, where the transformations are particular utility functions that also represent the individuals' risk attitudes. This class of theories satisfies Ex Ante Pareto, Dominance Package, Homogenous Anonymity, and Pareto Indifference (**Claim 1** in the Appendix). I say this is a class of theories, rather than a single theory, because different choices of privileged expected utility representations are possible.

An illustration may help us get a better grip on the Risk-Prioritarian approach, and also show

why some will find it unpalatable.[31] Suppose there are two individuals, and fix some profile in which the first individual is risk-neutral and the second risk-averse. Consider the following three pairs of choices in matrix form, where s^* is the disjunction of all states.

	s^*
a	(10, 0)
b	(0, 10)

	s^*
c	(5, 0)
d	(0, 5)

	s^*
e	(2, 0)
f	(0, 2)

Figure 2.4

It's easy to check that some versions of the Risk-Priority View deliver the following verdicts: relative to this risk profile, a is better than b, c and d are equally good, and f is better than e.[32] These Risk-Prioritarian theories give more weight to the first individual in the first choice and to the second individual in the third choice. But it's natural to think that people's risk attitudes shouldn't affect our judgments about the relative merits of a and b, and of e and f: intuitively, a and b are equally good, and so are e and f, since they distribute the same amounts of welfare, only to different people. Nor is this a peculiar feature of some Risk-Prioritarian theories; all versions of the Risk-Priority View will entail similar verdicts.[33]

Those tempted to reject this theory in light of these kinds of examples will want to embrace a different non-welfarist approach. In particular, we can consider a cousin of Homogenous Anonymity: the idea that, at least when there is no risk involved, equal quantities of welfare matter in the same way—independently of what risk attitudes people have.

> SURE-ACT ANONYMITY: For all sure acts $a, b \in A$, and for all profiles $R \in \mathcal{R}$, if a and b are permutations of one another, then neither $a \succ_R b$ nor $b \succ_R a$.

[31] This example is inspired by a case in Blessenohl (2020, p. 509). Blessenohl uses that case as an objection to a different theory of social choice, which incorporates Buchak's (2013)'s non-orthodox decision theory but shares with the Risk-Priority View a focus on giving weight to risk attitudes in situations where one might have thought them irrelevant.

[32] Suppose, for example, that $\mathbf{U}_{R_1}(10) = 10$, $\mathbf{U}_{R_1}(5) = 5$, $\mathbf{U}_{R_1}(2) = 2$, $\mathbf{U}_{R_1}(0) = 0$, and $\mathbf{U}_{R_2}(10) = 8$, $\mathbf{U}_{R_2}(5) = 5$, $\mathbf{U}_{R_2}(2) = 3$, $\mathbf{U}_{R_2}(0) = 0$.

[33] Incidentally, notice that the Risk-Priority View is sensitive to stakes: this means that, if we were working with multiple welfare profiles (as entertained in footnote 19 above), it would violate the relevant invariance condition, i.e., invariance with respect to a common positive affine transformation. I take this to be a feature of the view worth emphasizing, but not necessarily an objection—see Nebel's (2024) critical investigation of invariance principles, to which I owe this point.

Requiring Sure-Act Anonymity to hold, then, will serve as an antidote to the arguably objectionable verdicts of the Risk-Priority View considered above.[34]

2.5.2 The Unanimity Rule

Let us turn to examine a non-welfarist theory that respects Sure-Act Anonymity, as well as Ex Ante Pareto, Dominance Package, and Homogenous Anonymity (see **Claim 3** in the Appendix). This theory can be understood as a generalization of the idea that motivates Ex Ante Pareto. While Ex Ante Pareto says that autonomy matters, this theory says that it is all that matters. According to what we can call the Unanimity Rule:

- An act is at least as good as another just in case everyone weakly prefers it (adapted from Sen 2017, p. 75);
- An outcome is at least as good as another just in case the act that brings it about is weakly preferred by everyone.

This theory entertains much *incomparability* between acts and outcomes. If some individuals prefer a first act to a second, and others prefer the second to the first, then the two acts will be incomparable (in the sense that the social preference relation fails to hold either way). Similarly, two outcomes will be incomparable whenever there is no unanimity of preference over the respective acts that bring them about. Moreover, the Unanimity Rule is not well-defined unless we assume a tight restriction on which outcomes can result from which acts (**Claim 2** in the Appendix): roughly, what we need is that each outcome is the output of exactly one act—I call this the "Act-Outcome Restriction" in the Appendix. Recall that, in laying out my framework, I explicitly avoided assuming Rectangular Field, the claim that A includes all functions from states of nature to outcomes. We can now better see the reason for this choice.

[34]Thus, Sure-Act Anonymity is perhaps closer to the spirit of Chambers and Echenique (2012) and Fleurbaey and Zuber's (2017) principle than Independence of Irrelevant Alternatives—see footnote 27 above. As we will see in the next section, there are views that satisfy Sure-Act Anonymity while violating Independence of Irrelevant Alternatives and also Chambers and Echenique (2012) and Fleurbaey and Zuber's (2017) principle. These views do not intuitively take notice of risk attitudes in riskless contexts, suggesting that Sure-Act Anonymity is best suited among these principles to capture a concern with not letting risk attitudes make a difference when there is no risk.

A non-welfarist who cares about the value of autonomy may think that the history of how a certain social arrangement came about is ethically relevant. If so, outcomes in X will include such historical information, since outcomes include all normatively relevant information. But then we should want to impose strict requirements on which outcomes can be combined together, on pain of allowing in our model impossible acts: an example being an act a that delivers an outcome x in only some of the states, where part of the description of x is that it results from a context of zero risk.[35] Imposing these restrictions is part and parcel of the idea that motivates certain non-welfarist, autonomy-based approaches to social choice. The Unanimity Rule exemplifies this class of views: all that matters in evaluating an outcome, according to it, is what kind of act it resulted from. A rejection of Rectangular Field and the imposition of stringent requirements on what kinds of functions from states to outcomes constitute genuine acts is easily justified given this stance.[36]

This line of thought is bolstered by the following fact, proved in the Appendix: given Welfare-Risk Permissivism and Domain Condition*, a slight modification of the technical domain assumption required for Theorem 1, if Rectangular Field holds, no social quasi-ordering function satisfies Dominance Package, Ex Ante Pareto and Sure-Act Anonymity (**Claim 4** in the Appendix, which is essentially a variant of the result in Blessenohl 2020, though I draw a different lesson from it). Since these three latter principles can be motivated by a viable ethical outlook involving concern for autonomy, a certain kind of impartiality, and coherence in evaluating acts and outcomes, it seems entirely legitimate for those attracted to this outlook to take the fact just stated as an argument against Rectangular Field.

The real problem with the Unanimity Rule is that, while it fully accommodates autonomy, it gives no ethical weight at all to how well off people are. Suppose act a is preferred unanimously over b, and that everyone is equally well off in some possible outcome of a, x, as they are in some possible

[35] That a theory allows for such impossible entities is often seen in the literature as a strong objection against it; see, e.g., Broome (1991b, pp. 115-117) and Joyce (1999, p. 108).

[36] The Risk-Priority View, by contrast, is compatible with Rectangular Field because it does not need to include in an outcome's description any information about how it came about. Different non-welfarist theories accommodate Ex Ante Pareto in different ways—including historical information in an outcome is one such way, but it is not necessary.

outcome of b, y. Then the Unanimity Rule ranks x as better than y, violating Pareto Indifference. Worse still, suppose a involves some risk that everyone will end up badly off, in outcome z. The Unanimity Rule ranks z as better than any of the outcomes that b could result in, even if in some such outcomes everyone fares very well indeed. This is implausible; as I said at the beginning of the paper, everyone agrees that the well-being of individuals matters.

2.6 The Value of Welfare

How then can we accommodate the thought that welfare matters within a non-welfarist framework, so as to avoid the bad implication of the Unanimity Rule? The lesson seems to be that our theory should evaluate outcomes so that, if everyone is strictly worse off in x than in y, then x isn't ranked as socially better than y; in fact, we should want y to be ranked as better than x. Formally, the requirement our social quasi-ordering function would need to satisfy is

> EX POST PARETO. For all outcomes $x, y \in X$ and for all profiles $R \in \mathcal{R}$, if $W_i(y) > W_i(x)$ for all $i \in N$, then $y \succ_R x$.

Kaplow and Shavell (2001) show that given a widely accepted "continuity" assumption, any theory that violates the principle of Pareto Indifference entails violations of Ex Post Pareto, too. We can therefore understand Pareto Indifference as a condition that must be satisfied by theories wishing to avoid the sort of insensitivity to the value of welfare exemplified by the Unanimity Rule.

Here, however, we encounter a new problem, which serves as a counterpoint to Theorem 1. Let us step back: the Unanimity Rule stands as an alternative to the Risk-Priority View, in that it satisfies Sure-Act Anonymity. But, it turns out, it violates Pareto Indifference. The Risk-Priority View, by contrast, satisfies Pareto Indifference but violates Sure-Act Anonymity. The following theorem confirms that this is no coincidence.

> **Theorem 2.** *Given Welfare-Risk Permissivism and Domain Condition*, no social quasi-ordering function satisfies Dominance Package, Ex Ante Pareto, Sure-Act Anonymity, and Pareto Indifference.*

The proof can be found in the Appendix, and is based on a case due to Broome (1991b, p. 190) and Blessenohl (2020, p. 486); like Claim 4 above, Theorem 2 is closely related to Blessenohl's own result.[37] Domain Condition* is stated in the Appendix, too. The first main result of the paper was an impossibility involving welfarist theories: they are incompatible with Ex Ante Pareto, Dominance Package, and Homogenous Anonymity in my framework. This second result can be seen as an impossibility for non-welfarist theories. Even if we deny Independence of Irrelevant Alternatives, and so Welfarism, we need to choose between Pareto Indifference, which encodes a kind of concern for people's welfare, and Sure-Act Anonymity, which amounts to the thought that risk attitudes shouldn't affect what priority different people's welfare receives in non-risky choices.[38] More constructively, however, Theorem 2 helps us get a better view of the space of non-welfarist theories that satisfy Ex Ante Pareto, Dominance Package, and Homogenous Anonymity.

[37]The comparison is non-obvious because of differences in our respective frameworks. For one, Blessenohl's (2020) is a "single-profile" framework (individuals' risk attitudes are fixed) while mine is "multi-profile" (they vary with $R \in \mathcal{R}$). Further, Blessenohl bakes Pareto Indifference into that framework, and defines social preference over outcomes in terms of social preference over sure-acts. For reasons given above, if neither Pareto Indifference nor Rectangular Field are assumed, this definitional move is not available. Blessenohl shows that an analogue of Welfare-Risk Permissivism is inconsistent with the conjunction of statewise dominance and an analogue of Ex Ante Pareto, given the principle of "Constant Anonymity" (Blessenohl 2020, p. 495). Restated in my framework, Constant Anonymity says that for any profile $R \in \mathcal{R}$, and any two outcomes $x, y \in X$, if $W(x)$ and $W(y)$ are permutations of one another, then $x \sim_R^X y$. Note that Pareto Indifference follows from Constant Anonymity by taking the identity permutation. Given Rectangular Field, completeness of social preference, and Dominance Package, Sure-Act Anonymity and Constant Anonymity are equivalent. Therefore, under these assumptions Sure-Act Anonymity entails Pareto Indifference, and Theorem 2 adds nothing to Claim 4, which I called a variant of Blessenohl's result. Moreover, assume that for any list of welfare levels $(r_1, ..., r_n)$ such that $W(x) = (r_1, ..., r_n)$ for some $x \in X$, there exists a sure-act $a \in A$ whose unique outcome y is such that $W(y) = W(x) = (r_1, ..., r_n)$. Then given completeness of social preference and Dominance Package, the conjunction of Pareto Indifference and Sure-Act Anonymity is equivalent to Constant Anonymity, so again we see the connection. In my proof of Theorem 2, however, I assume neither Rectangular Field nor Pareto Indifference: instead, I use Dominance Package, Ex Ante Pareto, and Sure-Act Anonymity to derive a violation of Pareto Indifference. This re-framing of the significance of Blessenohl's case reflects my interest in seeing Theorem 2 as an exploration of (im)possibilities for non-welfarism.

[38]One might wonder whether Theorem 2 opens up an alternative way out of Theorem 1: if we deny Pareto Indifference to block this second impossibility, can we retain Independence of Irrelevant Alternatives after all? I haven't formally ruled this option out, but it is unpromising. The proof of Theorem 1 uses Pareto Indifference to derive a conclusion that is anyway surely correct for those who accept Sure-Act Anonymity; in the choice between c and d of Section 2.4, it is the claim that the outcome of c under E is neither better nor worse than the outcome of c under $\neg E$. This claim alone, coupled with Dominance Package and Ex Ante Pareto, allows us to derive violations of Independence of Irrelevant Alternatives.

2.7 Conclusion

As we conclude, it will be helpful to have a clear picture of the theoretical options that the work of this paper left open. Here it is.

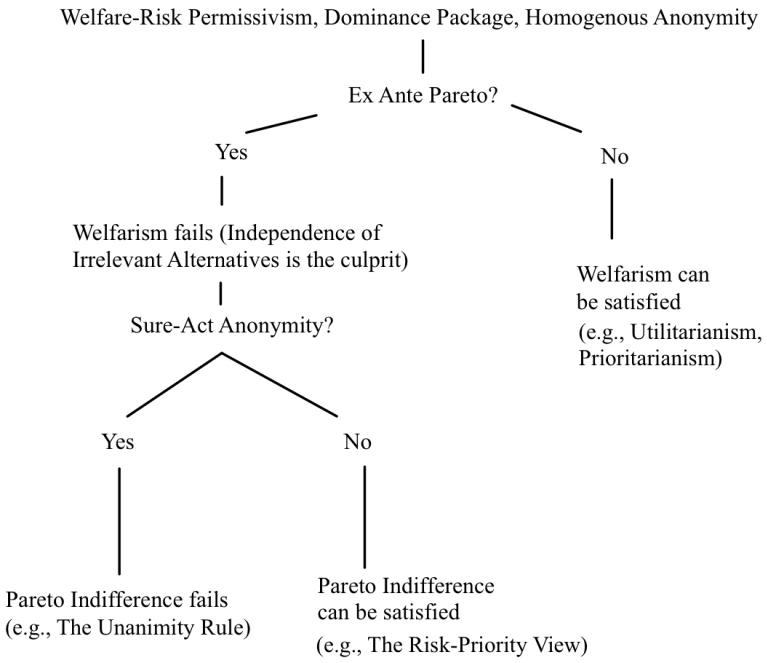


Figure 2.5

It is now clear that both welfarists and non-welfarists face difficult choices. To wrap up, let's begin by summarizing the welfarist's predicament. All of Welfare-Risk Permissivism, Dominance Package, and Homogenous Anonymity seem to me extremely plausible. But we should return in closing to what is perhaps the simplest way out for the welfarist: namely, to deny Welfare-Risk Permissivism. There are two ways to do this. One is to retain our assumption that there are independently given quantities of welfare, and insist that some particular attitude to risk involving one's welfare—for example, risk neutrality—is uniquely rational.[39] The other involves denying that quantities of welfare exist independently of choice under risk, in one of the two ways suggested in Section 2.2. Welfarists who find these claims hard to believe need to reject Ex Ante Pareto, and the value of autonomy it expresses. For those antecedently committed to either of these positions, the results

[39] See Zhao (2021) for a recent defence of risk neutrality, though note that Zhao's argument vindicates risk neutrality in a certain range of cases only (Zhao, 2021, p. 157; Wilkinson, 2022).

of the paper will look like additional evidence in their support.

But I'm inclined towards a different interpretation of the results: if both Welfare-Risk Permissivism and Ex Ante Pareto are true, somehow we need to thread the needle between the two possibilities that Theorem 2 leaves us with. Non-welfarist theories of social choice have been underexplored compared to their welfarist rivals, and much work remains to be done in figuring out what kind of non-welfarist one should be.[40] For those who reject Sure-Act Anonymity: can a plausible normative story be given for letting risk attitudes have normative relevance in non-risky choices? For those who reject Pareto Indifference: what alternative ways are there to accommodate the obvious fact that welfare matters? Without clear answers to these questions, our ability to simultaneously promote the values of welfare and autonomy remains at risk.[41]

Appendix

This appendix states some domain assumptions left out of the main text and gives statements and proofs of the Theorems and Claims mentioned therein. As a piece of convention applying throughout, for all $p, q \in \mathbb{R}$, let (p, q) denote:

- the n-tuple assigning p to the first $\frac{n}{2}$ entries and q to the remaining ones *if* n *is even*,
- the n-tuple assigning p to the first $\frac{n-1}{2}$ entries, q to the next $\frac{n-1}{2}$ entries, and some fixed arbitrary number $r \in R$ to the last entry *if* n *is odd*.

Recall that E and $\neg E$ partition S into two equiprobable events.

> DOMAIN CONDITION: For some $l, m, t \in \mathbb{R}$, with $l > m > t$ and $l - m = m - t$, there exist outcomes $v, w, x, y, z \in X$ and acts $a, b, c, d \in A : a(s) = v, \forall s \in S$, with

[40] However, see Chang (2001), Fleurbaey, Tungodden, and Chang (2003), Fleurbaey and Maniquet (2008), and the literatures on fairness and equality of opportunity mentioned above for discussions of extant non-welfarist approaches in social choice. Also relevant is my proposal in Cibinel (2022) responding to Blessenohl (2020), again cited above, which however needs to be supplemented to generate a full theory of social choice. In our taxonomy, any theory incorporating that proposal will fall in the bottom-left corner of the diagram sketched, together with the Unanimity Rule.

[41] As I see it, the lesson of the paper is in line with the program propounded by Sen (2017, pp. 339-44), of critically inspecting and, as appropriate, expanding the "informational basis" of social choice.

$W(v) = (l,t)$; $b(s) = w$, $\forall s \in S$, with $W(w) = (t,l)$; $c(s) = x$ for $s \in E$ and $c(s) = y$ for $s \in \neg E$, with $W(x) = (l,t)$ and $W(y) = (t,l)$; $d(s) = z, \forall s \in S$, with $W(z) = (m,m)$.

To illustrate in matrix form, Domain Condition simply guarantees the existence of the following acts, where, e.g., "$v : (l,t)$" indicates that the outcome is v and the list of welfare levels associated to v is (l,t).

	E	$\neg E$
a	v: (l,t)	v: (l,t)
b	w: (t,l)	w: (t,l)
c	$x : (l,t)$	$y : (t,l)$
d	$z : (m,m)$	$z : (m,m)$

Figure 2.6

For Theorem 2, we need the following variation on Domain Condition.

DOMAIN CONDITION*: For some $l, m, t \in \mathbb{R}$, with $l > m > t$ and $l - m = m - t$, there exist outcomes $q, r, u, v, w, x, y, z \in X$, and acts $a, b, c, d, e, f \in A$: $a(s) = x$ for $s \in E$ and $a(s) = y$ for $s \in \neg E$, with $W(x) = (l,m)$ and $W(y) = (t,m)$; $b(s) = v$ for $s \in E$ and $b(s) = w$ for $s \in \neg E$, with $W(v) = (m,l)$ and $W(w) = (m,t)$; $c(s) = z, \forall s \in S$, with $W(z) = (l,m)$; $d(s) = u, \forall s \in S$, with $W(u) = (m,l)$; $e(s) = q, \forall s \in S$, with $W(q) = (t,m)$; and $f(s) = r, \forall s \in S$, with $W(r) = (m,t)$.

In matrix form, this condition guarantees the existence of the following acts.

	E	$\neg E$
a	x: (l,m)	y: (t,m)
b	v: (m,l)	w: (m,t)
c	$z : (l,m)$	$z : (l,m)$
d	$u : (m,l)$	$u : (m,l)$
e	$q : (t,m)$	$q : (t,m)$
f	$r : (m,t)$	$r : (m,t)$

Figure 2.7

Theorem 1. *Given Welfare-Risk Permissivism and Domain Condition, no social quasi-ordering function satisfies Dominance Package, Ex Ante Pareto, Homogenous Anonymity, Pareto Indifference, and Independence of Irrelevant Alternatives.*

Proof. I show that if the first four principles are satisfied, Independence of Irrelevant Alternatives fails. Let $l, m, t \in \mathbb{R}$, $v, w, x, y, z \in X$, and $a, b, c, d \in A$ satisfy Domain Condition, so that $l > m > t$ and $l - m = m - t$; $a(s) = v$ for all $s \in S$, with $W(v) = (l, t)$; $b(s) = w$, with $W(w) = (t, l)$; $c(s) = x$ for $s \in E$ and $c(s) = y$ for $s \in \neg E$, with $W(x) = (l, t)$ and $W(y) = (t, l)$; and $d(s) = z$ for all $s \in S$, with $W(z) = (m, m)$.

Note that for all $i \in N$ with $c_i \neq d_i$, $\sum_{j=1}^{k} c_i(s_j)\pi(s_j) = \sum_{j=1}^{k} d_i(s_j)\pi(s_j)$ but c_i has uncertain consequences while d_i does not. By Welfare-Risk Permissivism and the fact that $\mathcal{R}$ is the set of all rational risk profiles, we can select $R \in \mathcal{R}$ such that $R_i = R_j$ for all $i, j \in N$, and $c_i P_i d_i$ for all $i \in N$ with $c_i \neq d_i$. And we can select $R' \in \mathcal{R}$ such that $R'_i = R'_j$ for all $i, j \in N$, and $d_i P'_i c_i$ for all $i \in N$ with $c_i \neq d_i$.

By Homogenous Anonymity, neither $a \succ_R b$ nor $b \succ_R a$. By Dominance Package (ii), neither $v \succ_R w$ nor $w \succ_R v$. By Pareto Indifference, $v \sim_R x$, $w \sim_R y$. By transitivity, neither $x \succ_R y$ nor $y \succ_R x$.

By Ex Ante Pareto, $c \succ_R d$ and $d \succ_{R'} c$. By Dominance Package (i) either $x \succ_R z$ or $y \succ_R z$, and either $z \succ_{R'} x$ or $z \succ_{R'} y$. There are four possibilities. If $x \succ_R z$ and $z \succ_{R'} x$, or $y \succ_R z$ and $z \succ_{R'} y$, we have a violation of Independence of Irrelevant Alternatives. Suppose next $x \succ_R z$ and $z \succ_{R'} y$. Since $x \succ_R z$ and $\neg(x \succ_R y)$, it follows by transitivity that $\neg(z \succ_R y)$. But then Independence of Irrelevant Alternatives is violated, given $z \succ_{R'} y$. Suppose finally $y \succ_R z$ and $z \succ_{R'} x$. Since $y \succ_R z$ and $\neg(y \succ_R x)$, it follows by transitivity that $\neg(z \succ_R x)$, which violates Independence of Irrelevant Alternatives given $z \succ_{R'} x$. □

Theorem 2. *Given Welfare-Risk Permissivism and Domain Condition*, no social quasi-ordering*

function satisfies Dominance Package, Ex Ante Pareto, Sure-Act Anonymity, and Pareto Indifference.

Proof. I show that if the first three principles are satisfied, Pareto Indifference fails. Let $l, m, t \in \mathbb{R}$, $u, v, w, x, y, z \in X$, and $a, b, c, d \in A$ satisfy Domain Condition*, so that $l > m > t$ and $l - m = m - t$; $a(s) = x$ for $s \in E$ and $a(s) = y$ for $s \in \neg E$, with $W(x) = (l, m)$ and $W(y) = (t, m)$; $b(s) = v$ for $s \in E$ and $b(s) = w$ for $s \in \neg E$, with $W(v) = (m, l)$ and $W(w) = (m, t)$; $c(s) = z, \forall s \in S$, with $W(z) = (l, m)$; $d(s) = u, \forall s \in S$, with $W(u) = (m, l)$.

Fix $R \in \mathcal{R}$ such that $a_i P_i b_i$ for all $i \in N$ such that $a_i \neq b_i$; by Welfare-Risk Permissivism and the fact that $\mathcal{R}$ is the set of all rational risk profiles, we can make this assumption. By Ex Ante Pareto, $a \succ_R b$. By Dominance Package (i), either $x \succ_R v$ or $y \succ_R w$.

Suppose that $x \succ_R v$. By Sure-Act Anonymity, neither $c \succ_R d$ nor $d \succ_R c$. By Dominance Package (ii), neither $z \succ_R u$ nor $u \succ_R z$. If $v \succ_R u$, or $u \succ_R v$, or neither $v \succeq_R u$ nor $v \succeq_R u$, Pareto Indifference is violated. So, suppose $v \sim_R u$. By transitivity, $x \succ_R u$. But given $\neg(z \succ_R u)$, it follows that $\neg(x \sim_R z)$, which violates Pareto Indifference. The case of $y \succ_R w$ is analogous, and requires a consideration of $e, f \in A$ as specified by Domain Condition*. □

Claim 1. *Risk-Prioritarian theories are social quasi-ordering functions that satisfy Ex Ante Pareto, Dominance Package, Homogenous Anonymity, and Pareto Indifference.*

Proof. It's easy to see that $\succeq_R^A$ and $\succeq_R^X$, as defined by the Risk-Priority View, are quasi-orderings. Now let $R \in \mathcal{R}$ be given. Choose $a, b \in A$ so that $a_i \neq b_i$ for some $i \in N$ and $a_i P_i b_i$ for all $i \in N$ such that $a_i \neq b_i$. Then $U_i(a_i) > U_i(b_i)$ for all i with $a_i \neq b_i$ and $U_i(a_i) = U_i(b_i)$ for all i with $a_i = b_i$. It follows that $U_R(a) > U_R(b)$. So Ex Ante Pareto is satisfied. For Dominance Package (i), suppose $a \succ_R b$. Then $U_R(a) > U_R(b)$, which cannot be true unless there exists $s \in S$ such that $U_R(a(s)) > U_R(b(s))$. For Dominance Package (ii), if $a(s) \succ_R b(s)$ for all $s \in S$, then $U_R(a(s)) > U_R(b(s))$ for all $s \in S$, and so of course $U_R(a) > U_R(b)$. For Pareto Indifference, if $W(x) = W(y)$ then $U_R(x) = U_R(y)$ and so $x \sim_R y$. Note that we are here using the assumption

that $W_i(x) \geq W_i(y)$ just in case $U_i(W_i(x)) \geq U_i(W_i(y))$, for all $i \in N$. Finally, if $U_i = U_j$ for all $i, j \in N$ and a and b are permutations of one another, then $U_R(a) = U_R(b)$. So, Homogenous Anonymity follows. □

We now give a formal statement of the Unanimity Rule. For all acts $a, b \in A$ and for all profiles $R \in \mathcal{R}$, define $a\overline{R}b$ to mean $a_i R_i b_i$ for all $i \in N$.

According to the Unanimity Rule:

- For all profiles $R \in \mathcal{R}$ and acts $a, b \in A$, $a \succeq_R^A b$ if and only if $a\overline{R}b$ (adapted from Sen 2017, p. 75);
- For all profiles $R \in \mathcal{R}$ and outcomes $x, y \in X$, $x \succeq_R^X y$ if and only if there exist $a, b \in A$, with $x = a(s)$ for some state $s \in S$ and $y = b(s)$ for some state $s \in S$, such that $a\overline{R}b$.

The Unanimity Rule requires the

> ACT-OUTCOME RESTRICTION. For all outcomes $x \in X$ and acts $a, b \in A$, if $x = a(s)$ for some state $s \in S$ and $x = b(s)$ for some state $s \in S$, then $a_i I_i b_i$ for all $i \in N$ and $R \in \mathcal{R}$; for all $x \in X$, there is some $a \in A$ such that $x = a(s)$ for some $s \in S$.

Claim 2. *If Act-Outcome Restriction fails, the Unanimity Rule contradicts the reflexivity of weak social preference.*

Proof. Consider the first conjunct of Act-Outcome Restriction, and suppose there exists $R \in \mathcal{R}$ such that $\neg(a\overline{R}b)$, with $x = a(s)$ for some $s \in S$, but also $x = b(s)$ for some $s \in S$. Then the Unanimity Rule entails $\neg(x \succeq_R x)$, contradicting reflexivity (moreover, if in addition $b\overline{R}a$, we can also deduce $x \succeq_R x$, with a contradiction following even apart from reflexivity). For the second conjunct, suppose there exists some $x \in X$ for which there is no $a \in A$ such that $a(s) = x$ for some $s \in S$. Then on the Unanimity Rule for all outcomes $y \in X$, $\neg(x \succeq_R y)$. So, in particular, $\neg(x \succeq_R x)$, again in violation of reflexivity. □

Claim 3. *Given Act-Outcome Restriction, the Unanimity Rule is a social quasi-ordering function that satisfies Ex Ante Pareto, Dominance Package, Homogenous Anonymity, and Sure-Act Anonymity.*

Proof. That the Unanimity Rule is a social quasi-ordering function given Act-Outcome Restriction follows from the fact that the individual preference relations over prospects $R_1, ..., R_n$ are all orderings. Ex Ante Pareto follows from the clause for $\succeq_R^A$, and Dominance Package by combination with the clause for $\succeq_R^X$. For Homogenous Anonymity, note that if a and b are permutations, and $R_i = R_j$ for all $i, j \in N$, either every individual is indifferent, or some prefer a and some prefer b. Either way, neither $a \succ_R b$ nor $b \succ_R a$. A similar argument establishes Sure-Act Anonymity. □

Claim 4. *Given Welfare-Risk Permissivism and Domain Condition*, if Rectangular Field holds, no social quasi-ordering function Satisfies Dominance Package, Ex Ante Pareto and Sure-Act Anonymity.*

Proof. We repeat the first paragraph of the proof of Theorem 2 (and so use Welfare-Risk Permissivism, Domain Condition*, Ex Ante Pareto, and Dominance Package (i)) to get $x \succ_R v$, with x, v and R defined as in that proof (note that we don't need the full strength of Domain Condition*). By Rectangular Field, we can find $c, d \in A : c(s) = x, \forall s \in S$ and $d(s) = v, \forall s \in S$. By Dominance Package (ii), since $x \succ_R v$, $c \succ_R d$, violating Sure-Act Anonymity. □

Chapter 3

The Risk-Priority View

Abstract

This paper develops and defends a neglected theory in distributive ethics, whose mark is a particular kind of sensitivity to people's risk attitudes. This theory, which I call the Risk-Priority View, evaluates distributions of well-being by first transforming each individual's well-being level in a certain way, and then summing these transformed well-being levels across individuals. Its distinctive feature is that the transformations are different for individuals with different attitudes to risk. I present plausible axioms that, in the presence of familiar background assumptions, turn out to be necessary and sufficient for the Risk-Priority View to hold. I then defend its striking verdict that people's risk attitudes affect the moral significance of their well-being even in the absence of any actual risk.

3.1 Introduction

How should we evaluate distributions of well-being across people? There are many options. The simplest approach is *Classical Utilitarianism*: the value of a distribution is given by summing each person's well-being. Another common approach is the *Priority View*, or *Prioritarianism*: the value of a distribution is given by first transforming each person's well-being in a certain way, and then summing such transformed well-being levels. The prioritarian transformation is the same for all individuals, and ends up giving more weight to worse off people. In this paper, I develop and defend a neglected theory for the evaluation of well-being distributions.[1] Like Prioritarianism, this

[1] I introduce this theory in **Chapter 2** of this dissertation (published as Cibinel 2025), but do not defend it there. Broome (1991b, pp. 211-213) hints at a similar view; see also Blackorby, Donaldson, and Weymark (1980, pp. 19-31), (2008, p. 159), and Dietrich (ms, section 5). Differences between my proposal and related literature are highlighted in footnotes below, as well as in the main text of Section 3.6.

theory evaluates each distribution of well-being in terms of a sum of transformed well-being levels. Unlike Prioritarianism, it does not perform the same transformation for all individuals. Instead, what transformation is performed on an individual's well-being level depends on that individual's preferences for risk. Specifically, according to what I will call the

> **Risk-Priority View**, social value is a sum of transformed well-being levels, where each individual's well-being level is transformed by a function that also represents this individual's risk attitude.

When people's risk attitudes differ, as they typically do, the Risk-Priority View entails that individuals should receive *differential treatment*: in some circumstances, we should prioritize the well-being of more risk-averse people; in others, the well-being of more risk-seeking individuals should be foregrounded instead.[2] Such verdicts contravene deeply held intuitions. It is standardly believed that

> (OUTCOME ANONYMITY) if the well-being distributions of two outcomes x and y only differ in who gets which well-being level, x and y are equally good (see, e.g., Adler 2019, p. 97),

and, even more strongly, that

> (SUPPES-SEN PRINCIPLE) if the well-being distributions of two outcomes x and y only differ in who gets which well-being level, and we improve the well-being of some individuals in x to get $x+$, then $x+$ is better than y (adapting Suppes 1966, Sen 2017, p. 211).

To think otherwise, it is typically said, amounts to being objectionably partial: to prioritize the interest of some individuals just because of their identity. However, this neglects another possibility—

[2]It is worth emphasizing from the outset that I have in mind the orthodox sense of risk attitudes with respect to some quantity—in our case, quantities of well-being: importantly, risk aversion and risk-seekingness in this sense *need not involve any departure from expected utility maximization*, which it will in fact prove necessary to use. See Arrow (1971) and Pratt (1964) for the orthodox treatment of risk attitudes. Buchak (2013) and Bottomley and Williamson (2024) provide rival treatments which do deviate from expected utility maximization. We will return to these non-orthodox approaches below.

that we deny these principles not because we think some individuals' interest matters more *simpliciter*, but rather because we think it matters more in the specific circumstances we find ourselves in. On the face of it, this attitude seems compatible with impartiality, for we may well think that the interest of those same individuals matters *less* in other circumstances, which would even things out. In a slogan: equal consideration for all individuals does not preclude differential treatment. This is exactly the way in which the Risk-Priority View rejects Outcome Anonymity and the Suppes-Sen Principle. While this kind of impartial differential treatment may initially appear odd, it is certainly not obviously reprehensible in the way its partial counterpart is. A central aim of the paper will be to provide reasons for thinking it may in fact be morally required.

The Risk-Priority View is closely related to John Harsanyi's (1955) famous social aggregation theorem, and to more recent contributions that build on Harsanyi's approach.[3] Harsanyi (1953), (1955), (1977), (1982) offered a refreshing perspective for thinking about how to evaluate distributions of well-being. The innovation was to look at how different theories about the social value of well-being distributions fare under *risk*, that is, in situations where we need to choose between policies that will result in different outcomes with different probabilities. My contribution sits squarely within this tradition, which has come to be known as the literature on social choice under risk.[4] I argue for the Risk-Priority View by showing that—given a proper understanding of the relationship between well-being and risk attitudes—it is uniquely positioned to accommodate two compelling principles for social evaluation in risky settings: that we should defer to rational people's assessment of their prospects under risk, and that the social value of policies should be assessed from an impartial point of view.

Here is an overview of what follows. Section 3.2 presents a pair of examples that illustrate how the Risk-Priority View works; these examples also preview a number of philosophical issues that the remainder of the paper then untangles. Section 3.3 lays out a framework for thinking about

[3] See, in particular, Blackorby, Donaldson, and Weymark (1980), (2008), Mongin (1994), Mongin and d'Aspremont (1998), Blackorby, Bossert, and Donaldson (2002), (2005, ch. 7), Fleurbaey and Zuber (2021a), Nebel (2022a), Dietrich (ms).

[4] Introductions to this literature are provided by Mongin and Pivato (2016), Fleurbaey (2018), and Adler (2019).

social choice under risk, and uses this framework to introduce the Risk-Priority View formally. Section 3.4 axiomatizes the class of Risk-Prioritarian theories—that is, of versions of the Risk-Priority View: I give necessary and sufficient conditions for a theory to count as Risk-Prioritarian. Since these conditions are very plausible, this result constitutes a powerful argument for the view. Section 3.5 defends the view's striking verdict that risk attitudes matter even in the absence of any actual risk, in a way that results in particularly counterintuitive violations of Outcome Anonymity and the Suppes-Sen Principle. I argue that such violations are supported by considerations of justifiability to each person and individual responsibility. Section 3.6 situates the proposal within related literature, highlighting connections and differences with extant approaches. Section 3.7 concludes the paper.

3.2 Priority by Risk Attitude

The aim of this section is to introduce the Risk-Priority View informally via examples, to see how it works in practice before we approach the topic with a higher degree of generality. The examples will also allow us to flag a range of philosophical issues surrounding the view, which will be tackled more thoroughly in later sections.

Very roughly, here is how the Risk-Priority View can be motivated. There are different ways to pursue one's good under risk: risk-averse individuals focus on avoiding worse case scenarios; risk-seeking individuals focus on pursuing better case scenarios. Accordingly, it seems somewhat worse for risk-averse people to end up very poorly off, compared to risk-seeking ones. And it seems to some extent better for risk-seeking people to end up very well off, compared to risk-averse ones. To illustrate, suppose there is a risk-averse individual and a more risk-prone one. And suppose further that we know one of them suffered a significant loss. On the line of thought I will pursue in this paper, it is worse if the risk-averse individual turns out to be the unlucky one, given the opportunities they are willing to sacrifice to avoid bad outcomes. Suppose next that one of them unexpectedly gained a fortune. On the view I will explore, it is better if the risk-seeking individual

turns out to be the lucky one, given the risks they are willing to incur to make top-quality outcomes more likely.

To unpack this thought, we need to draw conceptual distinctions between three notions at play in the examples just given. First, there are quantities of individual well-being: for example, an individual can suffer a *significant* loss in well-being. Second, there is a question of how much these quantities of well-being matter morally: for example, how bad it is that some given individual suffers a significant loss in well-being (Nebel 2022a, p. 39). Third, there is a measure of how risk-averse an individual is: according to orthodox decision theory, which the Risk-Priority View makes use of, this is given by an individual's "vNM utility" function (after von Neumann and Morgenstern 1944), representing the quantity that the individual maximizes the expectation of, under risk. Everyone should draw these distinctions conceptually; some, however, take a stance about their relationship.

Classical utilitarians say that the distinction between quantities of well-being and how much these quantities count is one without a difference (Broome 1991b). According to them, equal losses or improvements in well-being matter equally. So, the moral impact of a well-being loss or improvement just depends on how big it is. Prioritarians disagree: an equal improvement in well-being matters more, morally, the worse off an individual is (Parfit 2000; Adler 2011).

There is then the issue of how to relate an individual's well-being to her degree of risk aversion. Assuming the individual is making a self-interested decision, must her well-being be equated with her vNM utility? This goes at the heart of a longstanding debate about how to interpret standard decision theory (Fishburn 1989). Harsanyi thought the answer was "yes." Because of this, he interpreted his aggregation theorem—which proves that social value is a sum of individual vNM utilities—as an argument for Classical Utilitarianism (Harsanyi 1977, 1982). Many have objected to this interpretation on the grounds that it is a mistake to understand vNM utility as reflecting anything like strength of preference, or some other quantity that may reasonably be equated with

Figure 3.1

	Well-being	**Moral value**	**vNM utility**
Classical Utilitarianism	$=$	$=$	$\neq$
Harsanyi's Utilitarianism	$=$	$=$	$=$
Primitivist Prioritarianism	$\neq$	$\neq$	$\neq$
Technical Prioritarianism	$=$	$\neq$	$=$
The Risk-Priority View	$\neq$	$=$	$=$

the classical utilitarians' conception of well-being (Sen 1976; Weymark 1991). Prioritarians can also be sorted into how they address this issue: "technical" prioritarians equate well-being with vNM utility, "primitivist" prioritarians think of well-being as a separate quantity.[5]

The Risk-Priority View fills in a space that has so far been neglected: it distinguishes between quantities of well-being and of vNM utility, and says that the moral significance of well-being improvements and losses changes depending on the individual's vNM utility function, encoding their risk attitude. Figure 3.1 above offers a summary of the various options: the last row illustrates the Risk-Priority View, which takes well-being to be different from ($\neq$) vNM utility, and equates ($=$) the latter with moral value.[6]

With these distinctions in mind, we can characterize the earlier examples more precisely. Quantities of well-being are represented by numbers, and an individual's vNM utility encodes her risk attitude in relation to her well-being: the vNM utility difference that a given well-being improve-

[5] I am here following Greaves' (2015) terminology.

[6] Greaves (2015, p. 35) calls "modern utilitarianism" the view that vNM utility equals (what I have been calling) moral value. On this taxonomy, both Harsanyi's Utilitarianism and the Risk-Priority View count as versions of "modern utilitarianism." It seems to me that most modern utilitarians in this sense are either squarely Harsanyi-Utilitarians, or at any rate fail to take seriously the idea that there could be quantities of well-being independently of vNM utility (see Greaves 2017). A case that I find hard to categorize is Hammond (1996). His theory sums individual vNM utilities, where these are derived from ethical choices that bear upon each individual singly; unlike Harsanyi, he explains that these quantities are not what Classical Utilitarians have in mind, and rather have "[...] purely ethical significance" (p. 94). He nonetheless calls each individual's vNM utility so understood their "welfare" (ibid.). I am unsure how many "modern utilitarians" would side with the Risk-Priority View *if* they became convinced of its assumption that we can make sense of quantities of well-being independently of vNM utility, and how many would instead cling to the truly utilitarian aspect of their position and side with Classical Utilitarianism. As this paper makes clear, the two theories have very different motivations.

ment is mapped to measures the weight that that well-being improvement receives under risk. An individual who is risk-averse has a vNM utility function that is concave with respect to well-being, which means that as well-being levels get higher, further improvements in well-being of the same size are mapped to smaller vNM utility differences. If the individual is risk-seeking, her vNM utility function is convex, meaning that as well-being levels get higher, further improvements in well-being of the same size are mapped to larger vNM utility differences. For a risk-neutral individual, the vNM utility function is linear in well-being, so that all improvements of well-being of the same size are mapped to equal vNM utility differences. Consider now

> *Dire Circumstances.* Alda is risk-seeking and Bruno risk-averse. They are both in very dire circumstances, living miserably at well-being level 0. You can either (A) improve Alda's well-being to 0.5, or (B) improve Bruno's well-being to 0.5.

On Classical Utilitarianism, you should be indifferent between (A) and (B), since both actions improve the life of one individual by the same amount. On Prioritarianism, you should also be indifferent between (A) and (B), since both actions improve the life of one individual by the same amount, and both individuals are equally badly off to start with. In other words, both Classical Utilitarianism and Prioritarianism satisfy Outcome Anonymity. Moreover, suppose we consider improving Alda's well-being to 0.51 (A+). Then both Utilitarianism and Prioritarianism rank (A+) as better than (B), for they satisfy the Suppes-Sen principle.

But perhaps a case can be made that you should choose (B) in a comparison with (A), and even with (A+). Consider the disposition of risk-averse Bruno and of risk-seeking Alda. Bruno had made ample provision for securing a bare minimum level of well-being, at the cost of foreclosing various opportunities for good. Wouldn't it be especially bad if he ended up poorly off? Alda had knowingly factored in the possibility of failure, so as to improve chances at much better outcomes. Wouldn't it be more morally bearable if she were the one to end up poorly off? The Risk-Priority View is designed to capture this kind of verdict. Let $\mathbf{U}_A$ and $\mathbf{U}_B$ denote Alda's and Bruno's vNM utility functions respectively. For concreteness, we select the convex, and therefore risk-seeking, function $\mathbf{U}_A(x) = x^2$ for Alda and the concave, and therefore risk-averse, function $\mathbf{U}_B(x) = \sqrt{x}$

for Bruno. The view tells us to first transform their well-being levels by these vNM utility functions, and then sum such transformed well-being levels. So, we get

Value of action (A) $= \mathbf{U}_A(0.5) + \mathbf{U}_B(0) = (0.5)^2 + \sqrt{0} = 0.25$

Value of action (B) $= \mathbf{U}_A(0) + \mathbf{U}_B(0.5) = 0^2 + \sqrt{0.5} \approx 0.7$

Hence, on this version of the Risk-Priority View, you should choose (B) in a comparison with (A), violating Outcome Anonymity. It is easily checked that the value of action (A+) is about 0.26, so this version of the view enjoins you to choose (B) even in a comparison with (A+), violating the Suppes-Sen Principle. The view does not, however, unilaterally favor risk aversion. We can next examine

> *Favorable Circumstances.* Alda is risk-seeking and Bruno risk-averse. They are both in very favorable circumstances, living happily at well-being level 1. You can either (C) improve Alda's well-being to 1.5, or (D) improve Bruno's well-being to 1.5.

On both Classical Utilitarianism and Prioritarianism, you should be indifferent between (C) and (D), in line with Outcome Anonymity. Moreover, suppose we consider improving Bruno's well-being to 1.51 (D+). Then both Utilitarianism and Prioritarianism rank (D+) as better than (C), in line with the Suppes-Sen principle.

But perhaps a case can be made that you should choose (C) in a comparison with (D), and even with (D+). Consider again the disposition of risk-seeking Alda and of risk-averse Bruno. Alda is willing to incur serious risks so that she can improve chances at top-quality outcomes. Wouldn't it be especially good if she ended up at well-being level 1.5? Bruno, by contrast, is fine with giving up chances at the most valuable outcomes in order to avoid ending up badly off. Wouldn't it be satisfactory enough if he remained at well-being level 1? Again, the Risk-Priority View accommodates this kind of verdict. Employing the same vNM utility functions of the first example, we get

Value of action (C) $= \mathbf{U}_A(1.5) + \mathbf{U}_B(1) = (1.5)^2 + \sqrt{1} = 2.25$

Value of action (D) $= \mathbf{U}_A(1) + \mathbf{U}_B(1.5) = 1^2 + \sqrt{1.5} \approx 2.22$

So, on this version of the view, you should choose (C) in a comparison with (D), violating Outcome Anonymity. Since the value of action (D+) is about 2.23, this version of the view enjoins you to choose (C) even in a comparison with (D+), violating the Suppes-Sen Principle. As we will see below, different versions of the Risk-Priority View entail different specific verdicts. But, in general, the more risk-averse an individual is, the greater priority she gets from a social point of view when circumstances are dire. And the more risk-seeking an individual is, the greater priority she gets when circumstances are favorable.

3.3 Framework and Statement of the View

This section introduces a framework for social choice under risk that will allow us to state the Risk-Priority View more precisely; a formal presentation is given in the Appendix.[7] Well-being is assumed to be cardinally measurable and fully comparable. Informally, it makes sense to speak of *quantities* of well-being and to compare such quantities across people. There is a set of n individuals, and a set of final outcomes. A well-being function takes in outcomes and returns well-being distributions, understood as lists, with n entries, of numbers representing well-being levels on a scale encoding the assumptions about well-being just made: this is known as an *interval scale*, common for all individuals.

We then consider the set of probability functions on the set of outcomes, and interpret such functions as the policies or "lotteries" with possibly uncertain outcomes that we want to evaluate socially.[8] We can now treat any outcome x as equivalent to the "degenerate" probability function that assigns probability 1 to x and 0 to all other outcomes. This allows us to deal with a single object for social evaluation, namely, the set of lotteries (rather than two, the set of lotteries and the set of outcomes).

[7]The framework I use combines different extant approaches; see, in particular, Blackorby, Bossert, and Donaldson (2005, ch. 7), Blackorby, Donaldson, and Weymark (2008), Chambers and Echenique (2012), Fleurbaey and Zuber (2017), Sen (2017), Nebel (2024); I utilize a version of this framework in **Chapter 2** (Cibinel 2025).

[8]As clarified in the Appendix, we restrict our focus to all probability functions with finite support, i.e. which assign non-zero probability to finitely many outcomes.

Each lottery a gives to each individual i a prospect a_i: a probability function over well-being levels. We are interested in individuals' preferences over prospects. We assume that they are orderings—that is, reflexive, complete, and transitive relations. Moreover, we assume that when restricted to degenerate prospects, these orderings match (the ordering of) well-being levels: a degenerate prospect is preferred to another just in case it results in greater well-being.

We will be asking whether our evaluations of policies should change upon variations in people's preferences over prospects—that is, changes in their attitudes under risk. To address this, we need to consider different configurations of possible preferences over prospects that individuals might have. We will think of these as lists, with n entries, of orderings of prospects, and call them risk profiles, or simply profiles. The set of all risk profiles of interest will be denoted by $\mathcal{R}$. With $R = (R_1, ..., R_n) \in \mathcal{R}$, the i^{th} entry R_i stands in for individual i's ordering of prospects under that particular configuration. Finally, we are interested in evaluating what I will call *risk-encompassing social welfare functionals* (after Sen 2017's notion of a social welfare functional):

> A **risk-encompassing social welfare functional** is a function that returns a social ordering of lotteries for each risk profile of interest.[9]

This is the framework we shall work with, and the assumptions made so far are taken as given.

We next make extra rationality assumptions that are specific to the Risk-Priority View: in particular, that individual and social preference satisfy the axioms of expected utility theory. This assumption is now made without argument, but it will be justified by the axiomatization of Section 3.4, which will show that the key controversial axiom of expected utility theory—often called "independence"—follows from much weaker axioms in our setting. We accordingly assume that all entries of any risk profile in $\mathcal{R}$ satisfy the expected utility axioms and that the risk-encompassing social welfare functional outputs only social orderings that also satisfy the expected utility axioms. Individual and social preference orderings therefore admit of vNM utility representations. Our

[9]For similar objects of study, see Blackorby, Bossert, and Donaldson (2005, ch. 7), Blackorby, Donaldson, and Weymark (2008), Chambers and Echenique (2012), Fleurbaey and Zuber (2017), and **Chapter 2** (Cibinel 2025).

assumption that people's preferences for degenerate prospects match the ordering of well-being levels can now be stated as follows: individual vNM utility is a *strictly increasing* function of well-being.

We can finally present the Risk-Priority View. Informally, it works as follows. For each potential ordering of prospects an individual could have, the view establishes a particular way to weigh the well-being of individuals who end up with that risk attitude: a way to decide when to prioritize those individuals' interest *vis-à-vis* the interests of people with alternative preferences for risk. Given any assignment of risk attitudes to individuals, the view then evaluates an outcome by first transforming each individual's well-being in that outcome in the way established for that individual's risk attitude. Then, it takes the sum of these transformed well-being levels to deliver the overall social value of that outcome. Policies with uncertain outcomes are evaluated by taking the expectation of the sum of transformed well-being levels in each of the possible outcomes.

Formally, a risk-encompassing social welfare functional is Risk-Prioritarian if it evaluates policies in the following manner. Going through all rational orderings of prospects, $R_*, R_\dagger, R_\star, ...$, it picks out for us a *privileged* expected utility representation for each, $\mathbf{U}_{R_*}, \mathbf{U}_{R_\dagger}, \mathbf{U}_{R_\star}, ...$, respectively. Call this one-to-one correspondence between orderings and expected utility representations $\mathbf{U}$. Define next, for each profile R and lottery a,

(**Risk-Priority Formula for Policies**)

$$U_R(a) \coloneqq \mathbf{U}_{R_1}(a_1) + ... + \mathbf{U}_{R_n}(a_n),$$

where for each individual i, $\mathbf{U}_{R_i}$ is the privileged expected utility representation of R_i under $\mathbf{U}$, with R_i the ordering of individual i in profile R. The values assigned by $U_R(\cdot)$ to lotteries determine the social ordering under R according to our risk-encompassing social welfare functional: greater values mean better lotteries. When a lottery a assigns probability 1 to some outcome x, and so is equivalent to that outcome, we abuse notation and more conveniently write:

(Risk-Priority Formula for Outcomes)

$$U_R(x) \coloneqq \mathbf{U}_{R_1}(w_1) + ... + \mathbf{U}_{R_n}(w_n),$$

with $w_1, ..., w_n$ the well-being levels that $1, ..., n$ get in x.[10] This formulation better displays the key feature of the Risk-Priority View: when it comes to evaluating outcomes, each person's well-being is treated differently depending on their risk attitude.

The choice of *privileged* vNM utility functions encodes judgments about the relative merits of benefiting individuals with one risk attitude over individuals with some other risk attitude. To see the point, consider again the examples of Section 3.2, with Alda and Bruno. In the profile considered, Alda is risk-seeking and Bruno risk-averse. We can represent this graphically. Recall we picked a version of the Risk-Priority View where Alda's degree of risk-seekingness is represented by $\mathbf{U}_A(x) = x^2$ (unbroken blue line) and Bruno's degree of risk aversion by $\mathbf{U}_B(x) = \sqrt{x}$ (orange line; the dotted blue line will be mentioned momentarily).

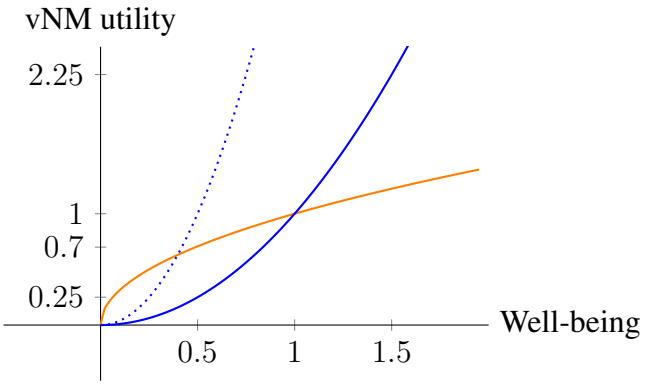


Figure 3.2

[10] Versions of this formula appear in Blackorby, Donaldson, and Weymark (1980, pp. 26, 28), (2008, p. 157), Broome (1991b, p. 211); compare Mongin and Pivato (2016, p. 723), Gustafsson, Spears, and Zuber (2023). Structurally, this is an additively separable representation: the term *generalized utilitarianism* is sometimes used for the special case where the transformations of well-being are the same (Blackorby, Bossert, and Donaldson 2002, p. 559), and *transformed utilitarianism* for the general case where they may vary (Blackorby, Donaldson, and Weymark 2008, p. 159). I call this the Risk-Priority formula because the transformations of well-being are to be understood specifically as vNM utility functions chosen in the impartial way defined by the Risk-Prioritarian method. None of the authors just cited defend the Risk-Prioritarian interpretation of the general formula as the correct way to rank outcomes, as is my goal here.

Well-being levels are plotted on the x-axis; the y-axis stands in for vNM utility units according to the privileged expected utility representations of our chosen version of the Risk-Priority View. The graph illustrates which individual we should prioritize in which circumstances: if we consider changing risk-averse Bruno's well-being from 0 to 0.5, we get approximately a 0.7 gain in vNM utility (orange line); for Alda, such a change in well-being occasions only a 0.25 gain in vNM utility (unbroken blue line). Hence the verdict that we should help Bruno in Dire Circumstances. For a change in well-being between 1 and 1.5, however, the difference in vNM utility for Bruno is only approximately 0.2, from 1 to about 1.2, while Alda's vNM utility gain is 1.25, from 1 to 2.25. Hence the verdict that we should help Alda in Favorable Circumstances.

Notice, however, that the function $\mathbf{U}'_A(x) = 4x^2$ (dotted blue line) is also a correct representation of Alda's degree of risk-seekingness. In general, vNM utility as a measure of an individual's degree of risk aversion is unique up to positive affine transformation: this means that we can transform any vNM utility function U representing an agent's risk attitude to $U' = \alpha U + \beta$, with $\alpha > 0$, and obtain another vNM representation of that individual's risk attitude. However, if we were to use $\mathbf{U}'_A(x) = 4x^2$ as our privileged representation of Alda's risk-seekingness, but kept $\mathbf{U}_B(x) = \sqrt{x}$ as our privileged representation of Bruno's risk aversion, we would get a different version of the Risk-Priority View. This version recommends prioritizing Alda over Bruno in both Dire and Favorable Circumstances, since Alda's change in vNM utility is greater than Bruno's, both when going from well-being level 0 to 0.5, and when going from well-being level 1 to 1.5.

The point here is that the Risk-Priority View is a *class* of theories, which differ from one another in the way they make interpersonal comparisons of vNM utility: that is, comparisons of the moral significance of well-being improvements and losses for people with different risk attitudes.[11] Some theories in this class are more plausible than others—for example, we would want the correct version not to prioritize Alda in both Dire and Favorable Circumstances. Particular normative

[11]The choice of privileged vNM utility functions is not fully unique: for $\alpha > 0$ and $\beta, \beta', \beta'', ... \in \mathbb{R}$, all and only one-to-one correspondences of the following form would generate the same social ordering of lotteries as $\mathbf{U}$, when plugged into the Risk-Prioritarian formula: R_* gets associated to $\alpha\mathbf{U}_{R_*} + \beta$, $R_\dagger$ gets associated to $\alpha\mathbf{U}_{R_\dagger} + \beta'$, $R_\star$ gets associated to $\alpha\mathbf{U}_{R_\star} + \beta'', ...$, with α constant.

judgments will have to guide us in making this choice.[12] There is an analogy here with Prioritarianism. The Priority View is also a class of theories, with each member of this class differing from the others in the degree to which it prioritizes worse off individuals. Some prioritarian theories implausibly give too much—or too little—weight to the worse off. Particular normative judgments will have to guide prioritarians in making their choice of which version of their theory to accept.

We have now presented the Risk-Priority View. What speaks in its favor? My first argument will be the axiomatization provided in the next section. I will then further motivate some of the most striking verdicts of the view by appealing to considerations of justifiability to each person and individual responsibility.

3.4 Axiomatization

The ethical axioms that I will introduce now are independently very appealing. In the presence of weak rationality assumptions on individual and social preferences, they characterize the class of Risk-Prioritarian theories. I give informal versions of the ethical axioms here, and formal versions in the Appendix. The first axiom is well-known: it's a variant of the "ex ante Pareto" principle, according to which we should defer to rational people's unanimous preference for a given policy over another. This principle encodes a certain kind of respect for rational people's evaluation of their prospects, in line with their risk attitudes. It can be stated precisely in the form of

> **STRONG PARETO**.
>
> (**Informal**) Consider any lotteries a, b and profile R. If every individual ranks their prospect with a at least as high as their prospect with b in R, then a is at least as good as b in R. If, in addition, some rank their prospect with a strictly higher than their prospect with b in R, then a is better than b in R.

The second axiom is new, and guarantees impartial consideration of all individuals, while allowing

[12] Those that follow Harsanyi need to make a parallel choice as to how interpersonal comparisons of vNM utility should be made—except that, for them, these are interpersonal comparisons *of well-being*. See Adler (2019, pp. 55-64) and Fleurbaey and Zuber (2021a). Section 3.6 further clarifies this difference.

that risk attitudes may be morally relevant. I will first illustrate how it works with an example. Suppose that there are three individuals, Cam, Dale, and Elise. We consider two risk profiles, R and R'. Under R, Cam is risk-averse, Dale risk-seeking and Elise risk-neutral. Under R', Cam and Dale swap risk attitudes: Cam is risk-seeking (to the same degree that Dale is in R), Dale is risk-averse (to the same degree that Cam is in R), and Elise is again risk-neutral. Consider now four lotteries a, b, c, d with sure outcomes, where $w^+ > w > w^-$ are numbers representing quantities of well-being.

	Cam	Dale	Elise
a	w	w	w^-
c	w^-	w^+	w^-
b	w	w	w^-
d	w^+	w^-	w^-

Figure 3.3

Notice that Elise is unaffected in a comparison between the four lotteries; so, we can forget about the last column in the table. Suppose c is socially better than a under R. This means that the gain of giving w^+ to risk-seeking Dale more than makes up for the loss of giving w^- to risk-averse Cam under R. R' differs only in that we swap Cam and Dale's risk attitudes. It stands to reason that the gain of giving w^+ to risk-seeking Cam should now, under R', outweigh the loss of giving w^- to risk-averse Dale, which amounts to d being better than b under R'. To think otherwise would effectively mean biasing one's social judgments in favor of Dale over Cam. Equal treatment of Cam and Dale, more broadly, requires a to be at least as good as c in R if and only if b is at least as good as d in R'.

To generalize this thought, we need to introduce two notions. First, an individual is "concerned" with respect to certain options just in case they potentially stand to gain or lose from a choice between those options. In the case just considered, Cam and Dale are concerned, while Elise is unconcerned. More precisely,

> if individual i's prospect with lottery a differs from at least one of their prospects with

lotteries b, c, d, then i is defined to be *concerned* with respect to lotteries a, b, c, d.

Second, a "permutation" σ of individuals is a one-to-one correspondence between individuals. If permutation σ assigns, for example, Dale to Cam, then we say that Dale is Cam's counterpart (relative to that permutation). We can finally introduce the novel axiom of

INTERPROFILE ANONYMITY.

(**Informal**) Consider any lotteries a, b, c, d and profiles R, R'. Suppose there is some permutation such that, for all individuals (i) the prospect they get with a equals the prospect their counterpart gets in b; (ii) the prospect they get with c equals the prospect their counterpart gets in d; and moreover, for all concerned individuals, (iii) the ordering they get with R equals the ordering their counterpart gets in R'. Then a is at least as good as c in profile R if and only if b is at least as good as d in profile R'.

To illustrate, consider again our example involving Cam, Dale, and Elise. With profiles R and R' as before, take the permutation σ of individuals defined as follows:

$\sigma(\text{Cam}) = \text{Dale},$

$\sigma(\text{Elise}) = \text{Elise},$

$\sigma(\text{Dale}) = \text{Cam}.$

For all individuals, the prospect they get with a (respectively, c) equals the prospect their counterpart gets with b (respectively, d). And for all concerned individuals, i.e. Cam and Dale, their ordering of prospects in R equals their counterpart's ordering of prospects in R'.[13] Interprofile Anonymity tells us that a is at least as good as c in profile R if and only if b is at least as good as d in profile R', as desired, capturing the attractive thought that the *identity* of who has a certain risk attitude should not be morally significant.

Note that Interprofile Anonymity is entailed by, and does not entail, the more standard "prospect anonymity" principle according to which if we swap prospects across individuals, our ranking of

[13] As it it happens, here this is also true for the unconcerned individual, Elise, but that is orthogonal to whether the antecedent of Interprofile Anonymity is satisfied.

lotteries should remain the same. Interprofile Anonymity says that if we swap prospects *and* swap risk attitudes *in the very same way* across individuals, our ranking should not change. If we think that the identity of individuals should not be morally relevant, but leave open that their risk attitudes may be, Interprofile Anonymity is the natural way to weaken the standard anonymity principle.

We can now state the formal result of the paper, which is proved in the Appendix at the end. The takeaway is that, under weak rationality assumptions for individual and social preference, Strong Pareto and Interprofile Anonymity characterize the class of Risk-Prioritarian theories. These weak rationality assumptions are stated in the Appendix, and I call their conjunction the Allais-Compatible Package, since it is consistent with the well-known "Allais preferences": a pattern of seemingly rational preferences that are nonetheless ruled out by expected utility theory, and which alternative decision theories have sought to accommodate. Our result nonetheless *derives* expected utility theory from the Allais-Compatible Package, Strong Pareto and Interprofile Anonymity; it therefore provides a reasonably strong justification for the Risk-Prioritarian reliance on expected utility theory.[14] Here is, then, a statement of the result.

> **Theorem.** *Assume the Allais-Compatible Package. Then a risk-encompassing social welfare functional satisfies Strong Pareto and Interprofile Anonymity if and only if it is Risk-Prioritarian.*

This theorem constitutes a powerful argument in favor of the Risk-Priority View. Strong Pareto displays a form of respect for people's autonomy under risk, while Interprofile Anonymity amounts to a very weak requirement of impartiality. The result above tells us that, in the presence of weak assumptions, only (and all) Risk-Prioritarian theories in distributive ethics can accommodate both. Examples of theories that violate Strong Pareto in this setting are Classical Utilitarianism and Prioritarianism (see **Chapter 2** (Cibinel 2025) for illustration). An example of a theory that

[14]The strategy for deriving expected utility theory is adapted from Mongin and Pivato (2015), McCarthy, Mikkola, and Thomas (2020), Nebel (2020) and Bradley (2022); in particular, the proof in the Appendix makes use of a case due to the latter two authors. In Cibinel (2022) and **Chapter 2** (Cibinel 2025), I explore what little room is left for Paretian theories that reject expected utility theory, and find that such theories face the challenge of properly accounting for the value of well-being (Cibinel 2025). For more discussion on deferring to the preferences of individuals that violate expected utility axioms, see Buchak (2017a), Blessenohl (2020), Nissan-Rozen (2021), and Thoma (2023).

violates Interprofile Anonymity would be one that systematically weighed one person's interests more heavily than another, irrespectively of what risk attitudes each individual ended up having.

3.5 The Risk-Priority *Ethos*

Given the framework of Section 3.3 and the theorem of Section 3.4, social evaluation that defers to rational people's assessment of their prospects, in line with Strong Pareto, *must* weigh people's well-being differently, according to their risk attitude, *even in risk-free cases*; Interprofile Anonymity then guarantees that such weightings are carried out impartially. The theorem of Section 3.4 does not, however, tell us *why* the Risk-Prioritarian ranking of outcomes is the morally correct one. Fundamentally, what motivates taking account of risk attitudes when there is no risk involved? Ostensibly, Strong Pareto only recommends giving weight to risk attitudes in risky contexts. The counterintuitive Risk-Prioritarian violations of Outcome Anonymity and the Suppes-Sen Principle, by contrast, result from giving weight to risk attitudes in risk-*free* contexts. What philosophical story could possibly get us from the former stance to the latter?

This section pursues two interrelated answers, concerning justifiability to each person and individual responsibility, respectively. To keep things simple, I will focus on violations of Outcome Anonymity; my remarks generalize to violations of the Suppes-Sen Principle. The strategy is this: I will introduce cases that proponents of Strong Pareto (or similar principles) already recognize as supporting some form of differential weighting of well-being, and describe how these verdicts are typically motivated. I will then show that such standard justifications of these verdicts should be developed and built upon in ways that lead to more radical violations of Outcome Anonymity, involving risk-free contexts.

3.5.1 Justifiability

Consider

Endorsed Risk, where risk-seeking Alda ends up poorly off at well-being level 0, as

a result of a risky policy that she wished the decision-maker would implement. Risk-averse Bruno lives a moderately good life at well-being level 0.5.

And compare

> *Unendorsed Risk*, where risk-averse Bruno ends up poorly off at well-being level 0, as a result of a risky policy that he wished the decision-maker would not implement. Risk-seeking Alda lives a moderately good life at well-being level 0.5

Intuitively, Alda's predicament in Endorsed Risk is less regrettable, morally, than Bruno's predicament in Unendorsed Risk. A natural explanation of this judgment appeals to the idea of justifiability to each individual affected (Scanlon 1998). In Endorsed Risk, Alda cannot reasonably blame the decision-maker for her unfortunate condition, since the decision-maker's choice perfectly reflected Alda's own evaluation of her prospects. When Alda ends up unlucky, the decision-maker can justify the decision to Alda on the basis of her preference *ex ante* (James 2012; Frick 2015). By contrast, Bruno appears to have grounds for complaint in Unendorsed Risk, since his unfortunate condition is due to the decision-maker enacting a policy that went against his own evaluation of his prospects.

These cases show how a commitment to respecting people's preferences *ex ante* in the way that Strong Pareto prescribes already entails some differential evaluation, resulting in violations of Outcome Anonymity (Cibinel 2022). In both Endorsed Risk and Unendorsed Risk, there is one individual at well-being level 0 and one at 0.5: yet, the outcome of Endorsed Risk is better than that of Unendorsed Risk.

It would be a further step, however, to argue for differential treatment in a context of zero risk: if we can benefit only one of Alda or Bruno under conditions of certainty, proponents of ex ante Pareto would typically say that we should be indifferent between the two options.[15] Consider, for example, the case of Dire Circumstances introduced in Section 3.2, where Alda and Bruno are in

[15] See, e.g., James (2012, p. 279), who would recommend flipping a fair coin, following Diamond (1967); I will momentarily operationalize indifference in precisely this way.

unfortunate conditions and you can either (A) improve Alda's well-being or (B) improve Bruno's well-being, but not both. If you choose (B), Alda is not badly off as a result of a gamble that she was willing to take, but rather because of a decision you took with full information of its consequences. Surely, one might think, in such a risk-free situation (A) and (B) are to be deemed equally good morally speaking, in line with Outcome Anonymity. By contrast, the Risk-Priority View recommends precisely taking the extra step of endorsing differential prioritization even in the absence of risk. In Dire Circumstances, for instance, some versions of the view deem (B) as better than (A).

I will now argue that these more radical violations of Outcome Anonymity can nonetheless be vindicated by considerations of justifiability to everyone involved. In other words, there is a viable route that takes us from the relatively uncontroversial

> **Ex Ante Intuition (I)**, according to which Alda ending up poorly off in Endorsed Risk is not as bad as Bruno ending up poorly off in Unendorsed Risk,

all the way to the verdicts of the Risk-Priority View in risk-free cases also: for example, the verdict that we should help Bruno rather than Alda in Dire Circumstances.

Notice, first, that whether an act is justifiable to every individual affected depends on the cumulative effects that can be expected, overall, if similar acts are performed in similar circumstances (Scanlon 1998, pp. 203-204). That is, the justifiability status of specific acts hinges on the justifiability status of principles "for the general regulation of behavior" (Scanlon 1998, p. 153)—in particular, for the regulation of behavior in related conditions. If a policy turns out to someone's disadvantage on some specific occasion, but is shown to be part of a broader scheme that can be expected to further this person's interest, it seems that the person has no reasonable ground for rejecting the specific act. In light of this clarification, we can ask: how would a general scheme of treating everyone equally in all circumstances, as Outcome Anonymity demands, fare against a Risk-Prioritarian plan of differential treatment?

To illustrate, suppose Alda, Bruno, and some decision-maker need to settle on a general princi-

ple for how the decision-maker will proceed should they find themselves in Dire Circumstances or in Favorable Circumstances, with no evidence making one scenario more likely than the other to obtain. On an Outcome Anonymity scheme, the decision-maker should be indifferent between whom to help in either circumstance. This scheme can be implemented along the following lines:

> (**S1**) toss a coin to decide between (A) and (B) if Dire Circumstances obtains, and toss a coin to decide between (C) and (D) if Favorable Circumstances obtains.

On a Risk-Prioritarian scheme, the decision-maker should prioritize risk-averse Bruno if circumstances are dire, and risk-seeking Alda if circumstances are favorable. That is, the decision-maker's scheme should be:

> (**S2**) choose (B) if Dire Circumstances obtains, and (C) if Favorable Circumstances obtains,

Which scheme would the group agree to enact?

Consider first S1. If Dire Circumstances obtains, the decision-maker will decide randomly who should be benefited, so Alda and Bruno have equal (and perfectly inversely correlated) chances of getting well-being level 0.5 or well-being level 0. If Favorable Circumstances obtains, the decision-maker will again decide randomly who should be benefited, so Alda and Bruno have equal (and perfectly inversely correlated) chances of getting well-being level 1 or well-being level 1.5. So, S1 looks exactly the same from Alda's perspective as it does from Bruno's. By contrast, S2 is more risky for Alda and more safe for Bruno. S2 doubles Alda's chance at getting well-being 1.5, since if the decision-maker is presented with Favorable Circumstances, they will benefit Alda for sure. However, S2 also doubles Alda's chance at well-being 0, since if the decision-maker is presented with Dire Circumstances, they will benefit Bruno for sure. Alda's chances of obtaining well-being 0.5 or 1 are correspondingly reduced to zero. By contrast, S2 doubles Bruno's chances at well-being levels 0.5 and 1, and reduces to zero his chance to obtain either well-being level 0 or 1.5.

Alda, being risk-seeking, will prefer S2 to S1: in expectation, the two schemes give her equal

future well-being, but the former increases her chance at the very best outcome. Bruno, being risk-averse, will also prefer that you adopt S2 instead of S1: in expectation, the two schemes give him equal future well-being, but the former reduces his chance at the worst outcome. Since S2 is unanimously preferred, it seems that the decision-maker can justify its adoption to all, while S1 can be reasonably rejected. In particular, the decision-maker can justify choosing (B) to Alda if Dire Circumstances in fact obtains, since it is part of a scheme that she has no reason to object to.[16] This example illustrates how, by construing risk-free choices as governed by general principles for how to choose in a *range* of interrelated scenarios, Risk-Prioritarian verdicts can be justified to every individual affected, along similar lines as the Ex Ante Intuition (I). In the long run, implementing a Risk-Prioritarian scheme instead of one conforming to Outcome Anonymity gives everyone more of what they want: more risks and opportunities to more risk-seeking individuals, and more safety and certainties to more risk-averse individuals. How could anyone object?

A ranking of policies in terms of their goodness from a social or moral point of view can be understood as a general principle specifying how a decision-maker should act in certain circumstances they might find themselves in. This suggests that a Risk-Prioritarian ranking can be justified to the individuals involved, whereas a ranking that respects Outcome Anonymity cannot. The route from Strong Pareto to Risk-Prioritarianism is not only logically necessitated, by the theorem of Section 3.4, but can also be motivated philosophically, by considerations of justifiability to each person.

3.5.2 Responsibility

Justifiability to each person is one of two main ways in which Strong Pareto and similar principles are typically motivated. The other, related approach appeals to individual responsibility.[17] Consider

Option & Brute Luck, where both Risk-seeking Alda and risk-averse Bruno end up

[16] Thomson (1990, p. 195) uses an argument with a similar structure to conclude that some versions of the trolley case are "[...] exceptions to the general rule that one may not kill one to save five."

[17] As noted by Fleurbaey (2010, p. 675), Hammond (1983) and Kolm (1998) are among those who connect respect for *ex ante* preferences to personal responsibility; I will discuss Dworkin (1981).

poorly off at well-being level 0. Alda's bad outcome, however, is the result of a gamble that she chose of her own will, whereas Bruno's bad outcome is the result of a gamble that he had no choice but to accept.

It seems right to say that Alda's predicament is less regrettable, from a moral point of view, than Bruno's predicament. Both events are bad, but Bruno's bad luck is worse. A natural explanation involves the thought that individuals should be held accountable for risks that they freely choose to incur (Dworkin 1981). When Alda ends up poorly off as a result of a risk that she decided to take, the thought goes, she is in some sense responsible for her unfortunate condition—this is a case of what Dworkin calls "option luck."[18] The same cannot be said of Bruno, who ends up badly off through no fault of his own—this is a case of what Dworkin calls "brute luck." Suppose we can either improve Alda's well-being level to 0.5, or else Bruno's well-being level to 0.5. Since Bruno's predicament is more regrettable, we should help Bruno instead of Alda, violating Outcome Anonymity.

When differential treatment is defended in this way, however, it is not extended to cover the more radical violations of Outcome Anonymity entertained by the Risk-Priority View, involving risk-free cases. In other words, there is no suggestion in this literature that we should prioritize Bruno over Alda in a case like Dire Circumstances, where neither Alda nor Bruno are in any obvious sense responsible for the unfortunate condition they find themselves in.[19] I will now argue that differential prioritization in risk-free cases is nonetheless justified by considerations broadly based on individual responsibility. In other words, there is a viable route that takes us from the

Ex Ante Intuition (II), according to which, in Option & Brute Luck, the fact that Alda ends up poorly off is not as bad as the fact that Bruno ends up poorly off,

all the way to the more controversial verdicts of the Risk-Priority View.

[18]The flipside of this is that if she turns out well off instead, she should be entitled to the outcome of her good luck. In this way, as Dworkin (1981) puts it, we can say to the risk-seeker that "[...] the possibility of loss was part of the life they chose—that it was the fair price of the possibility of gain" (p. 294). This passage is also cited by Fried (2003, p. 133), who offers an extensive critical discussion of Dworkin's position.

[19]Brute-luck egalitarians precisely want to neutralize the effects of such "brute luck" (see, e.g., Rakowski 1991, pp. 73-76—leaving aside the complication that Rakowski, like Dworkin and some other luck egalitarians, is interested in equality of resources and not of well-being).

To begin, it will help to think about how background risks and opportunities shape our judgments about responsibility for risks incurred. Suppose that, in

> *Background Risk*, there is a 50% chance that both Alda and Bruno will end up poorly off at well-being 0. Such a fact is given; there is nothing they can do about it. However, they can affect what happens if that bad outcome does not obtain. They can either have well-being level 0.5 in all remaining possibilities, or else well-being level 0 and well-being level 1.5 with equal probability. Risk-seeking Alda chooses the latter option, and risk-averse Bruno the former. As it happens, Alda ends up poorly off at well-being level 0.[20]

Is Alda here responsible for her unfortunate condition, in the way she is in Option & Brute Luck? I think we should feel ambivalent about how to answer this question, and that this supports a comparative understanding of individual responsibility for risks, which in turn counts in favor of the Risk-Priority View. Let's unpack this line of argument.

It might be tempting to say that whether Alda is responsible in Background Risk depends on whether she is at well-being level 0 because of the 50% background risk, or rather because of the added 25% risk she freely chose to incur. But this proposal does not stand up to scrutiny. If the mechanism that caused the background risk is exactly the same one that also caused the extra risk, it is not clear what it would mean for Alda to have ended up poorly off because of one or the other "portion" of the total risk she faced. It might be suggested that we need to know what would have happened had she chosen otherwise: if, had she chosen in the same way as Bruno, she would have ended up safe at well-being level 0.5, then she is responsible for her actual bad condition; but if, had she chosen in that way, she would still have ended up at well-being level 0, then she is not responsible for her actual bad condition.

[20] That background risks complicate Dworkin's distinction between option and brute luck is noted by Knight (2021) and Lippert-Rasmussen (2023). Neither of these authors, however, suggests that considerations of background risk point in the direction of the Risk-Priority View, which will be the main claim of this subsection.

But, first, there may be no fact of the matter as to what would have happened had Alda chosen otherwise (see Hare 2011, 2015). And, second, answering this question does not seem normatively relevant after all. Suppose the 50% background chance that Alda and Bruno end up at well-being level 0 in Background Risk is due to the presence of an incoming storm, which carries the threat of being hit by lightning. Each of them can stay at home: lightning will hit their house with 50% probability, but there is nowhere more protected where they can go. If their house remains intact, they will be safe but bored, at well-being level 0.5. Alternatively, they can go out surfing. If all goes well, they enjoy themselves, at well-being level 1.5; however, going surfing increases the chance of being hit by lightning to 75%. Alda goes surfing and gets hit by lighning. Suppose that Bruno, who stays home, is also hit by lightning. If we only have the resources to provide first-aid care to one of them, who should we help? I think the answer is that Bruno should be prioritized. And this is so even if we learn *ex post* that, had Alda remained at home, she would have been hit by lightning anyway. What matters to assessing Alda's situation, it seems to me, is that she chose to run more risk than Bruno did.[21]

The discussion does show, nonetheless, that the sense in which Alda is responsible for her bad condition in Background Risk is rather *sui generis*, and quite distinct from the more familiar notion of moral responsibility employed in attributions of praise and blame. Alda's odds were not ideal to start with, since she faced a 50% background risk that she could not get rid of. And it's not as if she wanted risk for its own sake: had she had the option of securing well-being level 1.5 with certainty—going surfing with a clear sky—she would have chosen that. Taking a gamble means that we are acting in sub-optimal circumstances, which must be taken into account when evaluating responsibility for risks incurred. This points to the suggestion that we interpret Alda's responsibility for her bad condition comparatively, to mean that she should receive less priority, in this case, than people in similar circumstances who took fewer risks, and that the reason for this is precisely that she chose to be more risk-seeking than those other people. That choice was hers,

[21] An analogy may be noted here with so-called "Frankfurt cases" in the literature on moral responsibility (Frankfurt 1969); however, I am about to sketch some differences between moral responsibility and the kind of responsibility for risks incurred of interest here.

and in this sense she is responsible for it. But this does not mean that, if we can help her at no cost to others, we have no reason to do so (cf. Lippert-Rasmussen 2023, sect. 7).

We can now notice, however, that this consideration also applies to Alda's situation in Option & Brute Luck: the reason why Alda took a gamble in that case must have been that her options were somewhat limited. If she could have acted so as to achieve the best possible outcome of that gamble with certainty, she would have done that. So, a comparative understanding of responsibility for risks incurred seems to be appropriate quite generally.

We next need to combine what we learned from Background Risk with the following realization: the choice of what risk attitude, or vNM utility function, to adopt is itself a *risky* choice, and can in fact be understood as the *only* risky choice we all face. To elaborate, let's begin by noting that preferences are dispositions to choose. Preferences under risk are therefore dispositions to choose when we don't know the consequences of our actions. They can thus be understood as complete plans for how to act in any given circumstance. Once I have chosen my risk attitude and I face some given decision problem, all I have to do is implement an algorithm: pick an option that maximizes my vNM utility. In making the "grand choice" of risk attitude—that is, in ranking all possible lotteries over well-being levels—each individual will take account of background risks and opportunities they might encounter: perhaps I will trip and break my neck; perhaps a million dollars will fall from the sky; and so on. Since we know about these risks and opportunities when we choose our risk attitude, they are relevantly like the 50% background risk that Alda faces in Background Risk. If she is responsible for her condition then—which she is, if we want to hold her responsible in Option & Brute Luck—so too are those of us who choose to be risk-seeking and end up poorly off "through no fault of their own."[22] If this seems odd to hear, we need to remember that the sense of personal responsibility at stake here is quite unlike the idea of moral responsibility associated with praise and blame, and simply provides a normative basis for our judgments of

[22]What of Dworkin's (1981) distinction between brute and option luck? My discussion can be interpreted as casting doubt on its usefulness (for a similar conclusion, see Bognar 2019); note that Dworkin himself points out that option and brute luck come in degree, and that the latter can be turned into the former through insurance (1981, p. 293). However, he hastens to add that this "[...] does not erase the distinction" (ibid.).

prioritization in cases where resources are limited. Perhaps surprisingly, then, Risk-Prioritarian verdicts are supported by considerations of personal responsibility similar to those that motivate the Ex Ante Intuition (II).

A Risk-Prioritarian ranking of policies in terms of how good they are socially or morally speaking systematizes such judgments of prioritization based on individual responsibility. Therefore, the route from Strong Pareto to Risk-Prioritarianism is not only logically necessitated, by the theorem of Section 3.4, but can also be motivated philosophically, by considerations of individual responsibility.

3.6 Comparison with Extant Work

To better appreciate what is novel about the Risk-Prioritarian proposal, we now relate it to some relevant extant literature, beginning with Harsanyi's (1955), (1977), (1982) social aggregation theorem. Such a comparison will allow us to pick back up on some themes introduced in Section 3.2. The aim of this section is to highlight key assumptions that differentiate the Risk-Priority View from extant approaches; it is not to convince someone already wedded to one of these alternatives to change their mind. Still, I will hint at reasons to side with the Risk-Prioritarian at each juncture. This section can be skipped without loss of continuity.

In Harsanyi's original set-up, there is no built-in assignment of well-being levels to outcomes, as given by our well-being function above. Rather, outcomes only include descriptive information about each individual's situation, such as what income this individual has. Moreover, individual preferences over prospects are fixed—it is a "single-profile" setting. Harsanyi shows that if (a single-profile version of) Strong Pareto is satisfied, and if individual and social preference satisfy the axioms of expected utility theory, then there exist expected utility representations $V, V_1, ..., V_n$ for society and individual preference respectively, such that for each lottery a,

$$V(a) = V_1(a_1) + ... + V_n(a_n)$$

(following Broome 2015's statement). When a lottery a assigns probability 1 to some specific outcome, a_i can be understood as a description of i's situation in that outcome—for example, what income i has. As already noted in Section 3.2, Harsanyi went on to interpret the individual expected utility representations in his formula as cardinal measures of individual well-being. He therefore took his theorem to support Classical Utilitarianism.

However, there are serious problems with Harsanyi's suggestion. A forceful critique has been mounted by Amartya Sen (1976), (1977), John Weymark (1991), and others (e.g., Roemer 1996). The complaint is that Classical Utilitarianism, as a position historically rooted in the tradition of Bentham, Mill, and Sidgwick, involves a cardinal notion of well-being that is prior to an individual's preferences in conditions of risk, as encoded by that individual's vNM utility. At best, Harsanyi hasn't given us sufficient reason to accept his interpretation of the V_i in the statement of his theorem as cardinal representations of well-being; at worst, it would be a mistake to so interpret them.[23] And, on reflection, there are good reasons to reject the vNM account of quantities of well-being (see, however, Broome 1991b, 2004, Jensen 1995, Risse 2002, McCarthy 2006, 2007, Greaves 2017, Adler 2019, for sophisticated variants of the opposing view). When philosophers tackle the issue of what well-being is, they ask the question:

> What would [...] make this person's life go, for him, as well as possible? (Parfit 1984, p. 493)

Overwhelmingly, the philosophical literature theorizes about this question *independently* of what risk attitude the given person has. This should come at no surprise: there are vastly more intuitive candidates for quantitative measures of well-being, such as intensity of happiness and suffering

[23]There are additional problems with Harsanyi's attempted derivation of Classical Utilitarianism, having to do with the fact that Harsanyi's theorem is single-profile. Mongin and d'Aspremont (1998) address these by reconstructing Harsanyi's aggregation theorem in Sen's (2017) "multi-profile" setting of social welfare functionals (following Mongin 1994). They then show (pp. 430-431) that an axiomatization of utilitarianism can be obtained by imposing standard axioms of anonymity ("Axiom A," p. 414) and independence of irrelevant alternatives ("Axiom I," p. 409). However, as they recognize (pp. 434-435), this move to a multi-profile setting does nothing to dispel the Sen-Weymark critique of using vNM utility functions as cardinal measures of individual well-being. As we shall presently see, the Risk-Priority View can be understood as the result of combining Mongin and d'Aspremont's rendering of Harsanyi's result in a multi-profile setting with the Sen-Weymark critique of Harsanyi's interpretation of his theorem. The role of their axioms A and I is played in my framework by Interprofile Anonymity, and the Sen-Weymark critique is accommodated by distinguishing between vNM utility and well-being.

(Kahneman, Wakker, and Sarin 1997), strength of preference elicited in contexts of certainty (Bell and Raiffa 1988), and metrics of human flourishing (VanderWeele 2017).[24]

It is here that the Risk-Priority View enters the stage. Recall that the framework of Section 3.2 assumed a cardinal notion of interpersonally comparable well-being that is independent of preferences under risk, thus addressing the Sen-Weymark critique head-on. The result of Section 3.4 can thus be interpreted as follows: having incorporated the Sen-Weymark observation, Harsanyi's theorem can be shown to support not Classical Utilitarianism, but rather the Risk-Priority View.[25]

On this picture, what is good for people does not depend on their preferences under risk, but what outcome is best from an impartial point of view for society at large does. In the course of arguing that there is an interpretation of Harsanyi's theorem that does not depend on assuming interpersonal comparisons of well-being, Nebel (2022a, p. 39) rightly points out that "[w]e should distinguish comparisons of different people's well-being from comparisons of the moral or social significance of their well-being." The Risk-Priority View *does* assume interpersonal comparisons of well-being; Nebel's lesson, however, can be applied all the same, in slightly modified form: we should distinguish between interpersonally comparable quantities of well-being and quantities of the "moral or social significance" of well-being. The Risk-Priority View takes the well-being function to measure the former, and the (privileged) individual vNM utilities to measure the latter.[26]

[24] **Chapter 1** presented my own preferred account. Dietrich (ms) shows how vNM utility can be decomposed into "pure risk attitude" and "pure utility" or well-being, lending further support to this approach. Fleurbaey and Zuber (2021a) follow the Harsanyi paradigm, whereby there is no antecedent cardinal measure of well-being, which is instead taken to be provided by vNM utility (see Fleurbaey and Zuber 2022, p. 463). However, at one point they claim that "[w]e could interpret income as a comprehensive and interpersonally comparable measure of ex post well-being [...]" (2021a, p. 378). This comment and the brief ensuing discussion (pp. 378-379) foreshadow a version of the approach I adopt here. See also their (2017) for an approach more in line with mine.

[25] Blackorby, Donaldson, and Weymark (2008, p. 157) incorporate the Sen-Weymark critique and derive profile-dependent transformed utilitarian social welfare functions—with the transformations given by vNM utility functions; however, they do not impose any axiom to ensure that the same transformation is applied to people with the same ordering of prospects, nor to relate different profile-dependent formulas in a normatively compelling way. Their (unrestricted) multi-profile result is instead an impossibility theorem (p. 161), as are those of Blackorby, Bossert, and Donaldson (2002, p. 587), (2005, p. 230). The Risk-Priority View escapes these impossibilities by denying welfarism (see **Chapter 2**, i.e. Cibinel 2025). The imposition of Interprofile Anonymity then guarantees impartial treatment of individuals, both within and across profiles.

[26] We can distinguish between the Risk-Priority View and a similar theory mentioned in passing by Broome (1991b, p. 211–213). On that theory (as on Broome's actual view) "the form of a person's [vNM] utility function represents the goodness or badness of risk to her" (p. 213); the primitive relation is betterness among prospects. On the Risk-Priority

To summarize, by the lights of a Risk-Prioritarian Harsanyi makes two mistakes: (i) he has an incorrect conception of well-being; (ii) he has an incorrect conception of the moral significance of well-being. In making these claims, the Risk-Prioritarian can rely on some fellow travelers, though none are companions for the whole ride: Classical Utilitarianism agrees with (i), but sides with Harsanyi on (ii). Primitivist Prioritarianism—the version of Prioritarianism which takes quantities of well-being as "primitive"—agrees with both (i) and (ii). As we have seen in Section 3.4, however, both Classical Utilitarianism and (both versions of) Prioritarianism violate Strong Pareto—the diagnosis for this lies in their refusal to equate vNM utility with moral value.

3.7 Conclusion

While weighing people's well-being by risk attitudes may seem odd at first, I have tried to convey a way of thinking about distributive ethics which makes the proposal quite natural. This approach is impartial in just the right way: enough as to not play favorites, but not so much that it disregards people's preferences. It can be motivated by interrelated considerations of justifiability to each person and individual responsibility. And it illuminates the philosophical significance of Harsanyi's social aggregation theorem. I conclude that the Risk-Priority View offers an attractive account for how well-being should be aggregated. After all, wouldn't it be a cosmic prank if the cautious among us ended up with either nothing or a lot, and those who dare led a so-so existence?

View, the form of a person's vNM utility function represents a person's preference for risk; the primitive relation is preference among prospects. It's plausible to think that facts about the goodness of prospects are necessary, and the same across people. By contrast, rational people can differ in their preferences for risk, which are contingent. As I see it (following Buchak and Nebel ms), the betterness relation among prospects is highly incomplete, with each completion of it a possible rational preference relation an individual can decide to adopt (see Gustafsson 2022 for arguments supporting the completeness of rational preferences). Given this picture, it is a distinctive feature of the Risk-Priority View that it places normative weight on people's *preferences*; the philosophical motivations developed in Section 3.5 support taking this route.

Appendix

This formal appendix lays out the model and axioms that the theorem is based upon, and then gives a proof of the theorem.

MODEL

• N is the finite set of n individuals.

• X is the set of final outcomes.

• $W : X \to \mathbb{R}^n$ is the well-being function.

ASSUMPTION 1. Well-being is cardinally measurable and fully comparable.

ASSUMPTION 2. The set of all well-being levels is order-isomorphic to $\mathbb{R}$.

ASSUMPTION 3. For any $(w_1, ..., w_n) \in \mathbb{R}^n$, there exists $x \in X$ such that $W(x) = (w_1, ..., w_n)$.

NOTATION. For all $i \in N$ and $x \in X$, $W_i(x)$ denotes i's well-being level w_i in x.

• A is the set of all probability functions on X with finite support.

NOTATION. For all $a \in A$, a_i is i's prospect with a.

NOTATION. A lottery that gives lottery a with probability α and lottery b with probability $(1 - \alpha)$ is written as $a\alpha b$.

• For all outcomes $x \in X$ and lotteries $a \in A$, if $a(x) = 1$, a and x are treated as equivalent, with corresponding abuse of notation.

• $\mathcal{R}$ is the set of all risk profiles of interest, i.e. n-tuples of rational orderings of prospects, with typical element $R = (R_1, ..., R_n)$.

NOTATION. For all $i \in N$ and $R \in \mathcal{R}$, strict preference is denoted by P_i and indifference by I_i.

• Risk-encompassing social welfare functionals are functions that return a social ordering of lotteries for each $R \in \mathcal{R}$, denoted by $\succeq_R$.

NOTATION. $\succ_R$ denotes strict preference, and $\sim_R$ indifference.

AXIOMS & LEMMAS

Rationality axioms.

Let $\succeq$ be any ordering on lotteries. Corresponding axioms are defined in the obvious way for any ordering on prospects. The following axioms are all well-known.

CONTINUITY. For all $a, b, c \in A$, if $a \succ b \succ c$, there exist $\alpha, \beta \in (0,1)$ such that $a\alpha c \succ b \succ a\beta c$.

For all outcomes $x \in X$ and lotteries $a \in A$, let $a[x]$ denote the probability, according to a, that an outcome at least as good as x obtains.

STOCHASTIC DOMINANCE. For all $a, b \in A$, if for all outcomes $x \in X$, $a[x] \geq b[x]$, then $a \succeq b$.

INDEPENDENCE. For all $a, b, c \in A$ and every $\alpha \in (0,1)$, if $a \succeq b$ then $a\alpha c \succeq b\alpha c$.

The Allais preferences (Allais 1953) are regimented as violations of T-Independence (Ozbek 2024, p. 5), which is entailed by Independence.

T-INDEPENDENCE. For all $a, b, c, d \in A$ and every $\alpha \in (0,1)$ if $a\alpha c \succeq b\alpha c$ then $a\alpha d \succeq b\alpha d$.

We next note two more entailments of Independence (adapting Ozbek 2024, pp. 5-6).

S-INDEPENDENCE. For all $a, b \in A$ and $\alpha \in (0,1)$, if $x_* \in X$ is such that for all outcomes y in the support of either a or b, $y \succeq x_*$, then $a \succeq b$ entails $a\alpha x_* \succeq b\alpha x_*$.

BETWEENNESS. For all $a, b \in A$ and $\alpha \in (0,1)$, if $a \succeq b$, then $a \succeq a\alpha b \succeq b$.

Lemma 1 (adapting Ozbek 2024, p. 6, Proposition 1). *T-Independence and S-Independence, and T-Independence and Betweenness, each entail Independence*.

Proof. The proof follows Ozbek (2024, pp. 16-17) with minor adaptations, and is omitted. □

We say that the **ALLAIS-COMPATIBLE PACKAGE** is satisfied if and only if: for all $R \in \mathcal{R}$, each individual entry of R satisfies Continuity and either S-Independence or Betweenness, and $\succeq_R$ satisfies Continuity and Stochastic Dominance.

Ethical axioms.

STRONG PARETO. For all profiles $R \in \mathcal{R}$ and lotteries $a, b \in A$, if $a_i R_i b_i$ for all i, then $a \succeq_R b$, and if, moreover, $a_i P_i b_i$ for some i, then $a \succ_R b$.

INTERPROFILE ANONYMITY. For all $a, b, c, d \in A$, and $R, R' \in \mathcal{R}$, if there exists a permutation σ of N such that for all $i \in N$, $a_i = b_{\sigma(i)}$ and $c_i = d_{\sigma(i)}$, and for all concerned individuals i, $R_i = R'_{\sigma(i)}$, then $a \succeq_R c \iff b \succeq_{R'} d$.

Lemma 2. *Strong Pareto, Interprofile Anonymity and the Allais-Compatible Package entail expected utility theory for individual and social preference.*

Proof. We derive Independence from the assumptions stated (adapting Nebel 2020 and Bradley 2022 to the framework of lotteries; note that we don't need Outcome Anonymity, which these authors make use of).

Suppose that individual preferences need not satisfy T-Independence. Then for some $R \in \mathcal{R}$, $R_i = R_j$ for all $i, j \in N$, and there exist prospects p, q, r, s and $\alpha \in (0, 1)$ such that $p\alpha s R_i q \alpha s$ but $q\alpha r P_i p \alpha r$ for all $i \in N$. Fix $a, b \in A$ as follows: $a = \underline{a}\alpha c$ and $b = \underline{b}\alpha c$, with everyone but two individuals, j and k, receiving some arbitrary well-being level w for sure, and $\underline{a}_j = p$, $\underline{a}_k = q$, $\underline{b}_j = q$ and $\underline{b}_k = p$, $c_j = s$, $c_k = r$. In particular, for all well-being levels w_p, w_q in the support of p and q respectively, choose $x \in X$ such that $W_j(x) = w_p, W_k(x) = w_q$, and let $\underline{a}(x) = p(w_p)q(w_q)$, and choose $y \in X$ such that $W_j(y) = w_q, W_k(y) = w_p$, and let $\underline{b}(y) = p(w_p)q(w_q)$. Note that there exists a bijection f from outcomes in the support of $\underline{a}$ to outcomes in the support of $\underline{b}$ such that $\underline{a}(x) = \underline{b}(f(x))$ and $W_j(x) = W_k(f(x))$, $W_k(x) = W_j(f(x))$; by Interprofile Anonymity, $x \sim_R f(x)$. Given this construction of a and b, Stochastic Dominance for social preference entails $a \sim_R b$. By Strong Pareto, however, $a \succ_R b$. Contradiction. Therefore, individual preferences satisfy T-Independence for all $R \in \mathcal{R}$, and hence Independence by Lemma 1. Since they also satisfy Continuity and are orderings, the vNM theorem applies and there exist expected utility representations for each entry R_i of any $R \in \mathcal{R}$ (see, e.g., Gilboa 2009, pp. 79-86). Moreover, since

individual preferences match well-being levels, each individual's vNM utility function is increasing in well-being.

Suppose next that social preference need not satisfy Independence. Then for some $R \in \mathcal{R}$ there exist $a, b, c \in A$ and $\alpha \in (0,1)$ such that $a \succeq_R b$ but $f = b\alpha c \succ_R e = a\alpha c$. Define a^*, b^*, c^* as follows: for all $i \in N$ a_i^* (respectively b_i^*, c_i^*) is the certainty equivalent of a_i (respectively b_i, c_i). By Strong Pareto $a^* \sim_R a, b^* \sim_R b$; hence $a^* \succeq_R b^*$. Let $e^* = a^*\alpha c^*$ and $f^* = b^*\alpha c^*$. By Stochastic Dominance for social preference, $e^* \succeq_R f^*$. By Strong Pareto, $e^* \sim_R e, f^* \sim_R f$; hence $e \succeq_R f$. Contradiction. Therefore, social preference satisfies Independence for all $R \in \mathcal{R}$. Since each output social preference is an ordering that also satisfies Continuity, the vNM theorem applies and for all $R \in \mathcal{R}$, there exists an expected utility representation for $\succeq_R$. □

THEOREM

Assume the Allais-Compatible Package. Then a risk-encompassing social welfare functional satisfies Strong Pareto and Interprofile Anonymity if and only if it is Risk-Prioritarian.

Proof. ($\Rightarrow$) Given Lemma 2, Harsanyi's (1955) social aggregation theorem allows us to infer that for all $R \in \mathcal{R}$, we can choose expected utility representations $U_R, U_{R_1}, ..., U_{R_n}$ for social and individual preferences respectively, such that for all $a \in A$, $U_R(a) = U_{R_1}(a_1) + ... + U_{R_n}(a_n)$ (see Broome 1991b for this way of stating Harsanyi's theorem). We next need to show the following:

(i) for all $R \in \mathcal{R}$, and $i, j \in N$, if $R_i = R_j$, then $U_{R_i} = U_{R_j} + \beta$ (this means: within a profile, the well-being of individuals with the same risk attitude is weighted in the same way);

(ii) for all $R, R' \in \mathcal{R}$, and $i, j, k, l \in N$, if $R_i = R'_j$ and $R_k = R'_l$ then $U_{R_i} = \alpha U_{R'_j} + \beta_1$ and $U_{R_k} = \alpha U_{R'_l} + \beta_2$, with constant $\alpha > 0$ and $\beta_1, \beta_2 \in \mathbb{R}$ (this means: across profiles, the well-being of individuals with the same risk attitudes is weighted in the same way).

We first prove that if (i) fails, then Interprofile Anonymity is violated. By contraposition, we

then get that if Inteprofile Anonymity holds, (i) also does. So, we fix $R \in \mathcal{R}$, and let $R_1 = R_2$ but $U_{R_1} = \alpha U_{R_2} + \beta$ for $1, 2 \in N$, with $\alpha \neq 1$. Without loss of generality, assume $\alpha > 1$. Then consider the following two lotteries $e, f \in A$, where $w^+ > w \in \mathbb{R}$.

	1	2	others
e	w^+	w	w
f	w	w^+	w

Figure 3.4

Lottery $e \in A$ results for sure in everyone but 1 at well-being w, and 1 at $w^+ > w$. Lottery $f \in A$ results for sure in everyone but 2 at well-being w, and 2 at w^+. Then $e \succ_R f$, which violates Interprofile Anonymity. To see this, let $R = R'$ and $a = d, b = c$ in the statement of the principle. Then take the permutation $\sigma(1) = 2, \sigma(2) = 1$, and $\sigma(k) = k$ for all $k \neq 1, 2 \in N$. For all individuals $i \in N$, we have $e_i = f_{\sigma(i)}$, and $f_i = e_{\sigma(i)}$, and for all concerned individuals, i.e. 1 and 2, we have $R_1 = R_{\sigma(1)=2}$ and $R_2 = R_{\sigma(2)=1}$. But it's not the case that $e \succeq_R f$ iff $f \succeq_R e$—since $e \succ_R f$. By contraposition, if Interprofile Anonymity holds, then (i) follows.

We show (ii) by contradiction. Fix $R, R' \in \mathcal{R}$, and $i, j, k, l \in N$ with $R_i = R'_j = \star$ and $R_k = R'_l = \circ$, but suppose for contradiction that $U_{R_i} = \alpha_1 U_{R'_j} + \beta_1$ and $U_{R_k} = \alpha_2 U_{R'_l} + \beta_2$, with $\alpha_1 \neq \alpha_2$. Without loss of generality, we assume that $\alpha_1 > \alpha_2$. Here is a depiction of R and R', and of another profile R'' that we will need to consider.

	i	j	k	l	others
R	$\star$	R_j	$\circ$	R_l	R_{others}
R'	R'_i	$\star$	R'_k	$\circ$	R'_{others}
R''	$\star$	R'_i	$\circ$	R'_k	R'_{others}

Figure 3.5

Arbitrarily select $w, w^+ \in \mathbb{R}$, where $w^+ > w$. Then, for $U_{R'_j}, U_{R'_l}, U_{R_i}, U_{R_k}$ the difference in utility between (degenerate prospects yielding for sure) w^+ and w is positive, since all individual utility functions are strictly increasing in well-being. Let ϵ be some positive real number strictly smaller

than any of these four differences in utility, and M some other positive real number strictly greater than any of these four differences in utility. Next, consider prospect v, which gives to the individual who receives it a p probability of w^+, and a $(1-p)$ probability of w, with $p = \frac{\epsilon}{2M}$. Notice that for all of $U_{R'_j}, U_{R'_l}, U_{R_i}, U_{R_k}$, the difference in utility between v and w is positive but strictly smaller than ϵ. This can be shown as follows, with U standing in for any one of the four utility functions under consideration. Since U is expectational,

$$U(v) = pU(w^+) + (1-p)U(w)$$

$$U(v) = pU(w^+) + U(w) - pU(w)$$

$$U(v) - U(w) = p[U(w^+) - U(w)]$$

$$U(v) - U(w) = \frac{\epsilon}{2M}[U(w^+) - U(w)] < \frac{\epsilon}{2M}M = \frac{\epsilon}{2} < \epsilon$$

Look next at the following table, illustrating four lotteries $e, f, g, h \in A$.

	i	j	k	l	others
e	v	w	w	w	w
f	w	v	w	w	w
g	w	w	v	w	w
h	w	w	w	v	w

Figure 3.6

Consider the permutation σ of N such that $\sigma(i) = j, \sigma(j) = i, \sigma(k) = l, \sigma(l) = k$, and $\sigma(m) = m$ for $m \neq i, j, k, l \in N$. Note that $g_i = h_{\sigma(i)}$ and $e_i = f_{\sigma(i)}$, and similarly for all other individuals. Also, $R''_i = R'_{\sigma(i)}$, and similarly for the remaining concerned individuals j, k, l. By Interprofile Anonymity, $g \succeq_{R''} e$ iff $h \succeq_{R'} f$. Now look at R and R'', and consider the identity permutation. By Interprofile Anonymity, $g \succeq_R e$ iff $g \succeq_{R''} e$. This can be seen by letting $a = b$ and $c = d$ in the statement of the principle. It then follows that $g \succeq_R e$ iff $h \succeq_{R'} f$. There are three cases. If $h \sim_{R'} f$, since $\alpha_1 > \alpha_2$, it follows tht $e \succ_R g$, which contradicts the verdict just reached using Interprofile Anonymity twice. If $f \succ_{R'} h$, we need an additional step. First notice that $f \succ_{R'} h$

entails that $U_{R'_j}(v) - U_{R'_j}(w) > U_{R'_l}(v) - U_{R'_l}(w)$. Define next v^+ as a prospect that gives to the individual who receives it a q probability of w^+ and a $(1-q)$ probability of w, with

$$q = \frac{U_{R'_j}(v) - U_{R'_j}(w)}{U_{R'_l}(w^+) - U_{R'_l}(w)}.$$

Notice that $0 < q < 1$, since $0 < U_{R'_j}(v) - U_{R'_j}(w) < \epsilon < U_{R'_l}(w^+) - U_{R'_l}(w)$. We now show that $U_{R'_l}(v^+) - U_{R'_l}(w) = U_{R'_j}(v) - U_{R'_j}(w)$. Since l is an expected utility maximizer,

$$U_{R'_l}(v^+) = qU_{R'_l}(w^+) + (1-q)U_{R'_l}(w)$$

Rearranging,

$$U_{R'_l}(v^+) - U_{R'_l}(w) = q[U_{R'_l}(w^+) - U_{R'_l}(w)]$$

So,

$$U_{R'_l}(v^+) - U_{R'_l}(w) = \frac{U_{R'_j}(v) - U_{R'_j}(w)}{U_{R'_l}(w^+) - U_{R'_l}(w)}[U_{R'_l}(w^+) - U_{R'_l}(w)] = U_{R'_j}(v) - U_{R'_j}(w)$$

With $g*, h* \in A$ as defined below, we have $h* \sim_{R'} f$ but $e \succ_R g*$ (since $\alpha_1 > \alpha_2$), which can be shown to contradict Interprofile Anonymity.

	i	j	k	l	others
e	v	w	w	w	w
f	w	v	w	w	w
$g*$	w	w	v^+	w	w
$h*$	w	w	w	v^+	w

Figure 3.7

To see the contradiction, consider again first profile R'' and R' as above, and note that Interprofile Anonymity entails $g* \succeq_{R''} e$ iff $h* \succeq_{R'} f$. Next, consider R'' and R, and note that Interprofile Anonymity entails $g* \succeq_R e$ iff $g* \succeq_{R''} e$. Therefore, $g* \succeq_R e$ iff $h* \succeq_{R'} f$. Finally, if $h \succ_{R'} f$, we need an analogous additional step. Define $v^\dagger$ as a prospect that gives to the individual who receives it an r probability of w^+ and a $(1-r)$ probability of w, with

$$r = \frac{U_{R'_l}(v) - U_{R'_l}(w)}{U_{R'_j}(w^+) - U_{R'_j}(w)}.$$

Notice that $0 < r < 1$. We show that $U_{R'_j}(v^\dagger) - U_{R'_j}(w) = U_{R'_l}(v) - U_{R'_l}(w)$. Since j is an expected utility maximizer,

$$U_{R'_j}(v^\dagger) = rU_{R'_j}(w^+) + (1-r)U_{R'_j}(w)$$

Rearranging,

$$U_{R'_j}(v^\dagger) - U_{R'_j}(w) = r[U_{R'_j}(w^+) - U_{R'_j}(w)]$$

So,

$$U_{R'_j}(v^\dagger) - U_{R'_j}(w) = \frac{U_{R'_l}(v) - U_{R'_l}(w)}{U_{R'_j}(w^+) - U_{R'_j}(w)}[U_{R'_j}(w^+) - U_{R'_j}(w)] = U_{R'_l}(v) - U_{R'_l}(w)$$

With $e*, f* \in A$ as given below, we have $h \sim_{R'} f*$ but $e* \succ_R g$, which again can be shown to contradict Interprofile Anonymity by similar reasoning as above.

	i	j	k	l	others
$e*$	$v^\dagger$	w	w	w	w
$f*$	w	$v^\dagger$	w	w	w
g	w	w	v	w	w
h	w	w	w	v	w

Figure 3.8

($\Leftarrow$) This direction is immediate, and the proof is omitted. □

Conclusion

The three chapters of this dissertation were written as stand-alone papers. Nonetheless, I hope the reader who ventured to look at all three was able to individuate the sketch of a general account for the evaluation of social outcomes and policies with uncertain consequences. For the sketch to turn into a full picture, many remaining questions would need to be addressed. As a way to conclude, I will bring up some that I find most interesting, or that trouble me. I will gesture at how I am inclined to respond, but an elaboration of the details is left for future work.

A first set of issues concerns the relationship between my preferred account of quantitative welfare comparisons, expounded in **Chapter 1**, and the vNM account, on which the vNM utility representation of a suitable preference or betterness relation over lotteries is taken as a measure of quantities of welfare. As has emerged at various points, **Chapter 2** and **Chapter 3** rely on a rejection of the vNM account. **Chapter 1** backs up this stance: if quantities of welfare can be derived independently of risk-preferences, it is implausible to think that vNM utility *also* happens to track quantities of welfare.

This thought, however, can be developed in a number of ways. First off, there are different versions of the vNM account itself. Harsanyi (1977), (1982) derives quantities of welfare (understood as preference intensities) from individual preferences over lotteries. Broome (1991b) derives quantities of personal good from a betterness relation over lotteries. Can different people's betterness relations over lotteries be different?[1] On an affirmative answer, the idea would be that a reasonably

[1]This question and the views that follow are taken up and explored in Buchak and Nebel (ms). As I noted in **Chapter 3**, footnote 26, I am sympathetic to a negative answer to the question, which is also the one Buchak and

risk-averse person's betterness relation will coincide with their reasonable preference relation. This would be different from the equally reasonable betterness relation of a risk-neutral person, which would match their risk-neutral preference relation. Alternatively, one might think that facts about the value of lotteries are invariant across people. In this case, welfare-risk permissivism can only be maintained by thinking of the betterness relation between lotteries as incomplete, with each completion of it a possible reasonable preference relation.[2] As we will now see, analogous choice points are available for the account of **Chapter 1**. The argument from welfare-risk permissivism to a denial of the vNM account works slightly differently depending on which versions of the two accounts we combine.

In **Chapter 1**, I assumed a single complete betterness relation over lives, and proceeded to derive quantities of welfare from ordinal comparisons. I left the issue of interpersonal comparisons of welfare aside, and effectively gave a treatment of quantities of welfare for a single individual. But if we are tempted to accept permissivism about risk attitudes, should we not also accept a form of permissivism about how the objective goods are weighed against one another? On this proposal, there are multiple equally reasonable ways to trade off objective goods. If we accept this claim, we can then say one of two things. Either different people's betterness relations over lives can be different, or betterness facts concerning lives are invariant across people. If individual betterness relations can differ, we should think of each person's betterness relation over lives as given by their reasonable preference relation. If betterness facts concerning lives are invariant across people, but we nonetheless want to say that different people can reasonably disagree over their tradeoffs between objective goods, then the betterness relation over lives should actually be understood as incomplete, *contra* the official position of all three chapters. Then, each completion of it is a possible reasonable preference relation.

Suppose we interpret both the vNM account and the Pluralistic Tradeoffs account as entertaining

Nebel favor.

[2] I am here glossing over the fact that welfare-risk permissivism was defined in **Chapter 2** in a state-contingent framework, as opposed to one of lotteries, to which we then switched in **Chapter 3**.

"betterness divergence" between people (a term due to Buchak and Nebel ms), or else as deriving quantities of welfare from reasonable preferences. Then the objection to the vNM account can be put like this: a person's preferences under risk need not be such that their vNM utility happens to track quantities of welfare for them, as given by their tradeoffs between objective goods. To impose such a constraint would amount to requiring risk neutrality. Suppose we interpret both the vNM account and the Pluralistic Tradeoffs account as entertaining incomplete betterness, and as deriving (correspondingly incomplete) quantitative welfare comparisons from it. Then the objection to the vNM account can be put like this: it would be a coincidence if the incomplete quantitative welfare comparisons entailed by the vNM account perfectly matched those entailed by the Pluralistic Tradeoffs account.

At present, however, I am sympathetic to a different position on the relationship between the two accounts. There is nothing *ad hoc* about accepting welfare-risk permissivism, but maintaining that the betterness relation over lives is complete and the same for all people. Welfare-risk permissivism is motivated by the idea that there should be some amount of discretion over how individuals evaluate the risks they face. Incompleteness or divergence in betterness over lives is similarly motivated by the idea that there should be some amount of discretion over how individuals evaluate their lives. But this latter thought can be accommodated by including desire or preference satisfaction as itself an objective good, along which people's lives can go better or worse (see Adler 2019, p. 71, fn. 22; Adler and Decanq 2022, p. 163). Once we do so, there seems to be no remaining motivation to accept incompleteness in ordinal welfare comparisons. We then have a single betterness relation over lives which applies to every person; yet people's preference relations over lotteries can diverge, and their betterness relations either also diverge or are incomplete. Thus, the vNM account fails.

I want to briefly mention an argument for the claim that, contrary to my approach, we should in fact expect the Pluralistic Tradeoffs account and the vNM account to agree on all their ver-

dicts.[3] For the purposes of explaining this argument, I assume betterness divergence for both the vNM account and the Pluralistic Tradeoffs account, and that quantities of welfare are derived from individual betterness relations (versions of this argument could be run for other combinations of the two views). As it happens, if we accept a seemingly compelling principle connecting objective goods and risk, a rapprochement between the two accounts will be *guaranteed*. Here is the principle.

> GOOD-BY-GOOD EX ANTE: If a lottery does at least as well as another with respect to all objective goods, it is at least as good; and if, moreover, it does better with respect to at least one good, then it is a better lottery.[4]

The reason for the equivalence of the two accounts, given Good-by-Good Ex Ante, is at heart just another application of what drives Harsanyi's aggregation theorem (Gorman 1968; Broome 1991b, ch. 4; Mongin and Pivato 2015).

Here is my reply. While Good-by-Good Ex Ante looks compelling, it is misleadingly so. There is in fact no notion of a lottery doing at least as well as another for a person with respect to a single good.[5] To see this, consider the four lotteries depicted below ($A, B, C,$ and D). We entertain only two objective goods, knowledge and friendship (columns), and two equiprobable states of the world (rows), which we think of as generating the lotteries. Level $k+$ (resp. $f+$) is a better level of knowledge (resp. friendship) than k (resp. f), which is in turn better than $k-$ (resp. $f-$). Suppose that life $(k+, f-)$ is equally good as life $(k-, f+)$ and also equally good as life (k, f). We can say these three lives instantiate the same welfare level, w. Suppose, moreover, that lives $(k+, f)$ and $(k, f+)$ are equally good as each other, instantiating welfare level $w+$; on the Pluralistic Tradeoffs account, this entails that lives $(k, f-)$ and $(k-, f)$ are also equally good as each other, instantiating

[3] I would like to thank Martin Peterson, Jeff Russell, and Ralph Wedgwood for prompting me to think about this objection to my approach.

[4] See Keeney and Raiffa's (1993, p. 230) additive independence; Weirich (2001, p. 155) talks of "BIT by BIT analysis" for a related idea (BIT stands for "basic intrinsic attitude").

[5] More precisely, there is no *interesting* notion of this kind: for example, if a lottery is sure to result in a better degree of a good than another, then we can derivatively say it is better for a person with respect to that single good. A reading of Good-by-Good Ex Ante which is equipped with this minimal notion is not enough to guarantee a rapprochement between the vNM and Pluralistic Tradeoffs accounts.

some lower welfare level $w-$.

A	Knowledge	Friendship
State 1	$k+$	$f-$
State 2	$k-$	$f+$

B	Knowledge	Friendship
State 1	k	$f-$
State 2	k	$f+$

C	Knowledge	Friendship
State 1	$k+$	f
State 2	$k-$	f

D	Knowledge	Friendship
State 1	k	f
State 2	k	f

Figure C.1

It is plausible to think that one could reasonably prefer A to B, and D to C. Lottery A and D result in welfare level w for sure. Lotteries B and C result with 50% chance in some higher welfare level $w+$, and with 50% chance in some lower welfare level $w-$. The Pluralistic Tradeoffs account entails that the welfare improvement from $w-$ to w is the same as that from w to $w+$. Someone who is risk-averse in welfare will then indeed prefer A to B and D to C.

But this pattern of preferences illustrates that we cannot assess a lottery by considering how it does along each objective good separately. If we look at knowledge only, A does exactly like C, giving $k+$ in State 1 and $k-$ in State 2. When this lottery on knowledge is paired with a lottery on friendship that gives $f-$ in State 1 and $f+$ in State 2, the overall lottery on welfare has no risk whatsoever. But when this lottery on knowledge is paired with a lottery on friendship that results in f for sure, the overall lottery in welfare *is* risky. If we look at knowledge only, B does exactly like D, resulting in k for sure. When this lottery on knowledge is paired with a lottery on friendship that gives $f-$ in State 1 and $f+$ in State 2, the overall lottery on welfare is risky. But when this lottery on knowledge is paired with a lottery on friendship that results in f for sure, the overall lottery in welfare has no risk. Whether a lottery on a single good is risky in welfare therefore depends on what lotteries on other goods it is paired with. What should matter to a person insofar as they are concerned with their interest is their welfare. Hence for a risk-averse individual the question of whether A is better than B as far as only knowledge is concerned suffers from presup-

position failure, as does Good-by-Good Ex Ante. There is no notion of a lottery being better than another for an individual as far as a single good is concerned. Indeed, this will be true for anyone who is not risk-neutral, which shows that we cannot use Good-by-Good Ex Ante to *argue* for a rapprochement between the vNM and the Pluralistic Tradeoffs accounts.

I would now like to point out some limitations of my statement and discussion of the Risk-Priority View (**Chapter 2** and **Chapter 3**). It is important to distinguish between two ways to understand the Risk-Priority View. Strictly speaking, the view is a theory for the social evaluation of outcomes and policies, where policies are understood in a technical sense as probability distributions over social outcomes. More broadly, however, one would hope for such a theory to substantially inform what an ethical deliberator should do in real-life choice situations.[6] I first mention two limitations of my statement of the Risk-Priority View *qua* theory for the social evaluation of social outcomes and lotteries, and then some limitations of the view *qua* guide to ethical decision-making.

Many of us are, no doubt, far from ideally rational. We are unlikely to have preferences that perfectly conform to the axioms of expected utility. But in that case, what is the moral value of our welfare, according to a proponent of the Risk-Priority View? The model introduced in **Chapter 2** and developed in **Chapter 3** is silent on this issue, since it assumes that the individuals involved are rational (though it is worth to reiterate that **Chapter 3** derives expected utility from weaker rationality conditions). One plausible answer is that, for each irrational person, we look for a preference relation that is rational and as close as possible to that person's actual preference relation. To be sure, it may be yet another idealization to think that we will be able to pick out a single rational substitute preference relation for every irrational person. There may be multiple equally plausible candidates. If so, this will introduce a certain amount of incompleteness in our expanded version of the Risk-Priority View: for each potential candidate preference relation, we run a version of the Risk-Priority View in which the person is associated with that preference relation, then take the

[6] In this spirit, **Chapter 3**, Section 3.2, used cases concerning what an agent should do to illustrate verdicts of the Risk-Priority View.

intersection of the resulting (complete) orderings.[7] To support this extension of the Risk-Priority View, we could appeal to resources used in **Chapter 3** to motivate some of its core verdicts. For example, it seems that an ethical deliberator could justify the use of a substitute rational preference relation in a manner that could not be reasonably rejected (Scanlon 1998). To an irrational person, they could plausibly say: "while I wanted to evaluate policies in a way that was most consistent with your preferences, the fact that they were not coherent threatened to impair my social evaluations. To remedy that, I evaluated policies for you departing the least I could from how you would have evaluated them yourself, while remaining coherent."

A second limitation of the Risk-Priority View as I formulated it concerns variable population cases. I only developed the view for a fixed number of individuals. But how should it be extended to evaluate policies and outcomes in which different numbers of people are affected? On its own, this question is unlikely to pose special difficulties. There are various ways to extend theories of distributive ethics from the fixed to the variable population case, and the available options for the Risk-Prioritarian will mirror those studied for additively separable theories in general (see, e.g., Blackorby, Bossert, and Donaldson 2005).[8]

Real issues are however brought up by variable population cases, once we consider one of their contexts of application. This is the ethics of future generations. The problem for the Risk-Prioritarian is that future people have no risk attitude in the present, since they do not yet exist. How should we then evaluate outcomes and policies that affect them?[9] What is challenging is not so much that future people do not yet exist, but rather that we do not know what risk attitude they

[7] See Sen (1973, ch. 3). The intersection $\succeq^I$ of two complete orderings $\succeq^1$ and $\succeq^2$ over the same domain is defined as follows: for all a, b in the domain, $a \succeq^I b$ if and only if both $a \succeq^1 b$ and $a \succeq^2 b$.

[8] The variable population context will also give the Risk-Prioritarian an opportunity to reiterate the importance of the distinction between interpersonal comparisons of well-being and of vNM utility (see **Chapter 3**). With variable populations, interpersonal *level* comparisons become key. Fleurbaey and Zuber (2021a, p. 383) propose a principle on which it is always better, morally speaking, for a population to be less rather than more risk-averse. This might be plausible if vNM utility is seen as a measure of well-being (as Fleurbaey and Zuber point out, after all, less risk-averse people have higher certainty equivalents). But if vNM utility is a measure of the moral value of well-being, this principle becomes open to question. If risk aversion does not impinge on an individual's well-being, why should the more risk-averse's well-being level always contribute less to the moral value of a population?

[9] See Buchak (2017b) for a take on this question different from the one I shall be gesturing at.

will have once they do exist. Many decision problems involving fixed populations will also face this difficulty, since it is unfeasible for policy-makers to poll every affected individual on their risk attitude. What needs to be emphasized is that we are now moving to address a concern about *how to use* the Risk-Priority View in practice: when an ethical deliberator lacks knowledge of people's risk attitude, how should they choose? It seems to me that there may not be any systematic answer (though see again Buchak 2017b for one such proposal). For example, if the ethical deliberator knows that the population is predominantly risk-averse, while being ignorant of specific individuals' risk attitudes, they will be justified in choosing predominantly risk-averse policies, and outcomes that favor equality in welfare. If they are completely in the dark about people's risk attitudes, many courses of action seem morally permissible; perhaps the deliberator's own risk attitude can here come into play.

Another practical issue concerns cases in which an ethical deliberator has evidence that differs from the evidence available to the affected individuals. Here the distinction between policies as lotteries over social outcomes and policies as acts that an agent may perform is crucial. The Risk-Priority View evaluates lotteries over social outcomes. If the individuals affected and the ethical deliberator all share the same evidence and probabilities, the agent's possible choices naturally correspond to lotteries over outcomes, with the probabilities used those shared by everyone. How a Risk-Prioritarian should act here is transparent. But if people's probabilities differ, or if the deliberator does not know what evidence the affected individuals have, matters complicate (this was touched upon in **Chapter 2**, Section 2.2—see again Broome 1987, p. 408). And here, too, there seems to be no general recipe. If the deliberator knows their evidence is strictly more informative than that available to the individuals affected, they seem justified in using it as their basis of assessment. But, in other cases, the deliberator's probabilities will play a less significant role in determining what they should choose. Suppose, for instance, that the deliberator knows the individuals affected all want a certain risky decision to be made, but does not know what evidence these individuals possess: the individuals may know more than the deliberator, or they may know less. At least in some versions of this situation, it seems that the deliberator should follow the

affected individuals' preference and choose in the way they all want.

This concludes my brief examination of some of the questions left unanswered in the three chapters. While it is no substitute for a full picture, it may hint at how I would go about developing one.

Bibliography

Adler, M. D. (2011). *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis*. Oxford University Press.

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

Adler, M. D. and Decancq, K. (2022). Well-being measurement. In Adler, M. D. and Norheim, O. F., editors, *Prioritarianism in Practice*, pages 128–171. Cambridge University Press.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole americaine. *Econometrica*, 21(4):503–546.

Arrow, K. J. (1951). *Social Choice and Individual Values*. Yale University Press (third edition, 2012).

Arrow, K. J. (1971). *Essays in the Theory of Risk-Bearing*. Chicago: Markham Publishing Co.

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

Bader, R. (2016). Conditions, modifiers, and holism. In Lord, E. and Maguire, B., editors, *Weighing Reasons*, pages 27–55. Oxford University Press.

Bader, R. M. (2018). Stochastic dominance and opaque sweetening. *Australasian Journal of Philosophy*, 96(3):498–507.

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

Blackorby, C., Bossert, W., and Donaldson, D. (2002). Utilitarianism and the theory of justice. In Arrow, K. J., Sen, A. K., and Suzumura, K., editors, *Handbook of Social Choice and Welfare: Volume 1*, pages 543–596. Elsevier.

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

Blackorby, C., Bossert, W., and Donaldson, D. (2006). Anonymous single-profile welfarism. *Social Choice and Welfare*, 27(2):279–287.

Blackorby, C., Donaldson, D., and Weymark, J. A. (1980). On John Harsanyi's defences of utilitarianism. *Discussion Paper of the Center for Operations Research and Econometrics, Université Catholique de Louvain*, 8013:1–40. Retrieved from https://john-weymark.github.io/files/BDW80CORE.pdf.

Blackorby, C., Donaldson, D., and Weymark, J. A. (2008). Social aggregation and the expected utility hypothesis. In Fleurbaey, M., Salles, M., and Weymark, J. A., editors, *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, pages 136–183. Cambridge University Press.

Blessenohl, S. (2020). Risk attitudes and social choice. *Ethics*, 130(4):485–513.

Bognar, G. (2019). Catering for responsibility: Brute luck, option luck, and the neutrality objection to luck egalitarianism. *Economics and Philosophy*, 35(2):259–281.

Bossert, W. and Weymark, J. A. (2004). Utility in social choice. In Barberà, S., Hammond, P. J., and Seidl, C., editors, *Handbook of Utility Theory: Volume 2*, pages 1099–1177. Springer.

Bottomley, C. and Williamson, T. L. (2024). Rational risk-aversion: Good things come to those who weight. *Philosophy and Phenomenological Research*, 108(3):697–725.

Bradley, R. (2022). Impartial evaluation under ambiguity. *Ethics*, 132(3):541–569.

Bradley, R. and Stefánsson, H. O. (2023). Fairness and risk attitudes. *Philosophical Studies*, 180:3179–3204.

Brenner, A., Maurin, A.-S., Skiles, A., Stenwall, R., and Thompson, N. (2021). Metaphysical explanation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition.

Broome, J. (1987). Utilitarianism and expected utility. *The Journal of Philosophy*, 84(8):405–422.

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

Broome, J. (1991b). *Weighing Goods: Equality, Uncertainty and Time*. Wiley-Blackwell.

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

Broome, J. (2015). General and personal good: Harsanyi's contribution to the theory of value. In Hirose, I. and Olson, J., editors, *The Oxford Handbook of Value Theory*, pages 249–266. Oxford University Press.

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

Buchak, L. (2017a). Taking risks behind the veil of ignorance. *Ethics*, 127(3):610–644.

Buchak, L. (2017b). Weighing the risks of climate change. *The Monist*, 102(1):66–83.

Buchak, L. and Nebel, J. M. (ms). The value of risky prospects. Retrieved from: personal communication.

Carlson, E. (2007). Higher values and non-archimedean additivity. *Theoria*, 73(1):3–27.

Carlson, E. (2015). Organic unities. In Hirose, I. and Olson, J., editors, *The Oxford Handbook of Value Theory*, pages 285–299. Oxford University Press.

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

Chang, H. F. (2001). Non-welfarist paretian methods of policy assessment. Retrieved from: http://dx.doi.org/10.2139/ssrn.288085.

Chateauneuf, A. and Wakker, P. (1993). From local to global additive representation. *Journal of Mathematical Economics*, 22(6):523–545.

Cibinel, P. (2022). Risk attitudes and justifiability to each. *Ethics*, 133(1):106–121.

Cibinel, P. (2025). Welfare and autonomy under risk. *Philosophy and Phenomenological Research*, 110(2):526–551.

Cohen, H., Maril, A., Bleicher, S., and Nissan-Rozen, I. (2022). Attitudes toward risk are complicated: Experimental evidence for the re-individuation approach to risk-attitudes. *Philosophical Studies*, 179:2553–2577.

Crisp, R. (2000). Particularizing particularism. In Hooker, B. and Little, M. O., editors, *Moral particularism*, pages 23–47. Oxford University Press.

Danan, E., Gajdos, T., and Tallon, J.-M. (2015). Harsanyi's aggregation theorem with incomplete preferences. *American Economic Journal: Microeconomics*, 7(1):61–69.

Dancy, J. (2004). *Ethics Without Principles*. Oxford University Press.

Darwall, S. (2002). *Welfare and Rational Care*. Princeton University Press.

Debreu, G. (1960). Topological methods in cardinal utility theory. In Arrow, K. J., Karlin, S., and Suppes, P., editors, *Mathematical Methods in the Social Sciences, 1959. Proceedings of the First Stanford Symposium*, pages 16–26. Stanford University Press.

Diamond, P. A. (1967). Cardinal welfare, individualistic ethics, and interpersonal comparison of utility: Comment. *Journal of Political Economy*, 75(5):765–766.

Dietrich, F. (ms). Welfare vs. utility. Retrieved from: https://philpapers.org/rec/DIEWVU.

Doody, R. (ms). Consider the ostrich: Non-utilitarians, ex ante interests, and burying your head in the sand. Retrieved from: http://rdoody.com/ConsiderOstrich.pdf.

Dreier, J. (1996). Rational preference: Decision theory as a theory of practical rationality. *Theory and Decision*, 40(3):249–276.

Dworkin, R. (1981). What is equality? Part 2: Equality of resources. *Philosophy and Public Affairs*, 10(4):283–345.

Dyer, J. S. and Sarin, R. K. (1979). Measurable multiattribute value functions. *Operations Research*, 27(4):810–822.

Ferreira, F. H. G. and Peragine, V. (2016). Individual Responsibility and Equality of Opportunity. In Adler, M. D. and Fleurbaey, M., editors, *The Oxford Handbook of Well-Being and Public Policy*, pages 746–784. Oxford University Press.

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

Fletcher, G. (2013). A fresh start for the objective-list theory of well-being. *Utilitas*, 25(2):206–220.

Fletcher, G. (2015). Objective list theories. In Fletcher, G., editor, *The Routledge Handbook of Philosophy of Well-Being*, pages 148–160. Routledge.

Fletcher, G. (2016). *The Philosophy of Well-Being: An Introduction*. Routledge.

Fleurbaey, M. (2009). Two variants of Harsanyi's aggregation theorem. *Economics Letters*, 105(3):300–302.

Fleurbaey, M. (2010). Assessing risky social situations. *Journal of Political Economy*, 118(4):649–680.

Fleurbaey, M. (2018). Welfare economics, risk and uncertainty. *Canadian Journal of Economics/Revue canadienne d'économique*, 51(1):5–40.

Fleurbaey, M. and Maniquet, F. (2008). Utilitarianism versus fairness in welfare economics. In Fleurbaey, M., Salles, M., and Weymark, J. A., editors, *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, pages 263–280. Cambridge University Press.

Fleurbaey, M., Tungodden, B., and Chang, H. F. (2003). Any Non-welfarist Method of Policy Assessment Violates the Pareto Principle: A Comment. *Journal of Political Economy*, 111(6):1382–1385.

Fleurbaey, M. and Voorhoeve, A. (2013). Decide as you would with full information! An argument against ex ante Pareto. In Norheim, O., Hurst, S., Eyal, N., and Wikler, D., editors, *Inequalities in Health: Concepts, Measures, and Ethics*, pages 113–128. Oxford University Press.

Fleurbaey, M. and Zuber, S. (2017). Fair management of social risk. *Journal of Economic Theory*, 169:666–706.

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

Fleurbaey, M. and Zuber, S. (2021b). Universal social welfare orderings and risk. Retrieved from: https://dx.doi.org/10.2139/ssrn.3876726.

Fleurbaey, M. and Zuber, S. (2022). A new puzzle in the social evaluation of risk. *Economics and Philosophy*, 38(3):450–465.

Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy*, 66(23):829–839.

Frick, J. (2013). Uncertainty and justifiability to each person: Response to Fleurbaey and Voorhoeve. In Norheim, O., Hurst, S., Eyal, N., and Wikler, D., editors, *Inequalities in Health: Concepts, Measures, and Ethics*, pages 129–146. Oxford University Press.

Frick, J. (2015). Contractualism and social risk. *Philosophy and Public Affairs*, 43(3):175–223.

Fried, B. H. (2003). Ex ante/ex post. *Journal of Law and Contemporary Issues*, 13:123–160.

Friedman, M. (1953). Choice, chance, and the personal distribution of income. *Journal of Political Economy*, 61(4):277–290.

Gilboa, I. (2009). *Theory of Decision under Uncertainty*. Cambridge University Press.

Goldschmidt, Z. and Nissan-Rozen, I. (2021). The intrinsic value of risky prospects. *Synthese*, 198:7553–7575.

Gonzales, C. (1996). Additive utilities when some components are solvable and others are not. *Journal of Mathematical Psychology*, 40(2):141–151.

Gonzales, C. (2000). Two factor additive conjoint measurement with one solvable component. *Journal of Mathematical Psychology*, 44(2):285–309.

Gonzales, C. (2003). Additive utility without restricted solvability on every component. *Journal of Mathematical Psychology*, 47(1):47–65.

Gorman, W. M. (1968). The Structure of Utility Functions. *The Review of Economic Studies*, 35(4):367–390.

Greaves, H. (2015). Antiprioritarianism. *Utilitas*, 27(1):1–42.

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

Greco, J. (2012). A (different) virtue epistemology. *Philosophy and Phenomenological Research*, 85(1):1–26.

Griffin, J. (1986). *Well-Being: Its Meaning, Measurement, and Moral Importance*. Oxford: Clarendon Press.

Gustafsson, J. E. (2022). *Money-Pump Arguments*. Cambridge University Press.

Gustafsson, J. E., Spears, D., and Zuber, S. (2023). Utilitarianism is Implied by Social and Individual Dominance. IZA Discussion Paper 16561, IZA Institute of Labor Economics. Retrieved from: https://docs.iza.org/dp16561.pdf.

Hájek, A. (2021). Risky business. *Philosophical Perspectives*, 35(1):189–205.

Hammond, P. (1983). Ex-post optimality as a dynamically consistent objective for collective choice under uncertainty. In Pattanaik, P. K. and Salles, M., editors, *Social Choice and Welfare*, volume 145 of *Contributions to Economic Analysis*, pages 175–205. Elsevier.

Hammond, P. J. (1981). Ex-ante and ex-post welfare optimality under uncertainty. *Economica*, 48:235–250.

Hammond, P. J. (1982). Utilitarianism, uncertainty and information. In Sen, A. K. and Williams, B. A. O., editors, *Utilitarianism and Beyond*, pages 85–102. Cambridge University Press.

Hammond, P. J. (1996). Consequentialist decision theory and utilitarian ethics. In Farina, F., Hahn, F., and Vannucci, S., editors, *Ethics, Rationality, and Economic Behaviour*, pages 92–118. Oxford University Press.

Hare, C. (2010). Take the sugar. *Analysis*, 70(2):237–247.

Hare, C. (2011). Obligation and regret when there is no fact of the matter about what would have happened if you had not done what you did. *Noûs*, 45(1):190–206.

Hare, C. (2015). Statistical People and Counterfactual Indeterminacy. In Eyal, N., Cohen, I. G., and Daniels, N., editors, *Identified versus Statistical Lives: An Interdisciplinary Perspective*, pages 124–136. Oxford University Press.

Harsanyi, J. C. (1953). Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy*, 61(5):434–435.

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

Harsanyi, J. C. (1977). *Rational Behaviour and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press.

Harsanyi, J. C. (1982). Morality and the theory of rational behaviour. In Sen, A. and Williams, B., editors, *Utilitarianism and Beyond*, pages 39–62. Cambridge University Press.

Hedden, B. and Muñoz, D. (2024). Dimensions of value. *Noûs*, 58(2):291–305.

Hedden, B. and Nebel, J. M. (2024). Multidimensional concepts and disparate scale types. *Philosophical Review*, 133(3):265–308.

Hooker, B. (2000). Moral particularism: Wrong and bad. In Hooker, B. and Little, M. O., editors, *Moral particularism*, pages 1–22. Oxford University Press.

Hurley, S. (1989). *Natural Reasons: Personality and Polity*. Oxford University Press.

Jaffray, J.-Y. (1974). On the extension of additive utilities to infinite sets. *Journal of Mathematical Psychology*, 11(4):431–452.

James, A. (2012). Contractualism's (not so) slippery slope. *Legal Theory*, 18(3):263–292.

Jensen, K. K. (1995). Measuring the size of a benefit and its moral weight. On the significance of John Broome's interpersonal addition theorem. *Theoria*, 61(1):25–60.

Joyce, J. M. (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press.

Kahneman, D., Wakker, P. P., and Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *The Quarterly Journal of Economics*, 112(2):375–405.

Kaplow, L. and Shavell, S. (2001). Any non-welfarist method of policy assessment violates the Pareto principle. *Journal of Political Economy*, 109(2):281–286.

Keeney, R. L. and Raiffa, H. (1993). *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press.

Keller, S. (2009). Welfarism. *Philosophy Compass*, 4(1):82–95.

Kiesewetter, B. (2017). *The Normativity of Rationality*. Oxford University Press.

Knight, C. (2021). An argument for all-luck egalitarianism. *Philosophy and Public Affairs*, 49(4):350–378.

Kolm, S.-C. (1998). Chance and justice: Social policies and the Harsanyi–Vickrey–Rawls problem. *European Economic Review*, 42(8):1393–1416.

Kowalczyk, K. (2023). People in suitcases. *Journal of Moral Philosophy*, 20:3–30.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (1971). *Foundations of Measurement. Volume I: Additive and Polynomial Representations*. Dover. Unabridged republication, 2007; original publication by Academic Press.

Kreps, D. (1988). *Notes on the Theory of Choice*. Routledge.

Köbberling, V. and Wakker, P. P. (2003). Preference foundations for nonexpected utility: A generalized and simplified technique. *Mathematics of Operations Research*, 28(3):395–423.

Lederman, H. (forthcoming). Of marbles and matchsticks. *Oxford Studies in Epistemology*, 8. Retrieved from: https://philpapers.org/archive/LEDOMA-2.pdf.

Lippert-Rasmussen, K. (2023). Justice and bad luck. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2023 edition.

Lord, E. (2018). *The Importance of Being Rational*. Oxford University Press.

Luce, R. D. and Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1(1):1–27.

McCarthy, D. (2006). Utilitarianism and prioritarianism I. *Economics and Philosophy*, 22(3):335–363.

McCarthy, D. (2007). Measuring life's goodness. *Philosophical Books*, 48(4):303–319.

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

Mogensen, A. L. (2024). Respect for others' risk attitudes and the long-run future. *Noûs*, 58(4):1017–1031.

Mongin, P. (1994). Harsanyi's aggregation theorem: multi-profile version and unsettled questions. *Social Choice and Welfare*, 11:331–354.

Mongin, P. (2016). Spurious unanimity and the Pareto principle. *Economics and Philosophy*, 32(3):511–532.

Mongin, P. and d'Aspremont, C. (1998). Utility theory and ethics. In Barberà, S., Hammond, P. J., and Seidl, C., editors, *Handbook of Utility Theory: Volume 1*, pages 371–481. Kluwer Academic Publishers.

Mongin, P. and Pivato, M. (2015). Ranking multidimensional alternatives and uncertain prospects. *Journal of Economic Theory*, 157:146–171.

Mongin, P. and Pivato, M. (2016). Social evaluation under risk and uncertainty. In Adler, M. and Fleurbaey, M., editors, *The Oxford Handbook of Well-Being and Public Policy*, pages 711–745. Oxford University Press.

Moore, A. and Crisp, R. (1996). Welfarism in moral theory. *Australasian Journal of Philosophy*, 74(4):598–613.

Moore, G. E. (1903). *Principia Ethica*. Cambridge University Press.

Nebel, J. M. (2020). Rank-weighted utilitarianism and the veil of ignorance. *Ethics*, 131(1):87–106.

Nebel, J. M. (2022a). Aggregation without interpersonal comparisons of well-being. *Philosophy and Phenomenological Research*, 105(1):18–41.

Nebel, J. M. (2022b). Totalism without repugnance. In McMahan, J., Campbell, T., Ramakrishnan, K., and Goodrich, J., editors, *Ethics and Existence: The Legacy of Derek Parfit*, pages 200–231. Oxford University Press.

Nebel, J. M. (2023). The sum of well-being. *Mind*, 132(528):1074–1104.

Nebel, J. M. (2024). Ethics without numbers. *Philosophy and Phenomenological Research*, 108(2):289–319.

Nissan-Rozen, I. (2017). How to be an ex-post egalitarian and an ex-ante Paretian. *Analysis*, 77(3):550–558.

Nissan-Rozen, I. (2021). Weighing and aggregating reasons under uncertainty: A trilemma. *Philosophical Studies*, 178:2853–2871.

Nozick, R. (1974). *Anarchy, State, and Utopia*. Basic Books.

Oddie, G. (2001). Axiological atomism. *Australasian Journal of Philosophy*, 79(3):313–332.

Ozbek, K. (2024). Expected utility, independence, and continuity. *Theory and Decision*, 97(1):1–22.

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

Parfit, D. (2000). Equality or priority? In Clayton, M. and Williams, A., editors, *The Ideal of Equality*, pages 81–125. New York: St. Martin's Press.

Parfit, D. (2011). *On What Matters. Volume One*. Oxford University Press.

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica*, 32(1/2):122–136.

Rakowski, E. (1991). *Equal Justice*. Oxford University Press.

Raz, J. (2000). The truth in particularism. In Hooker, B. and Little, M. O., editors, *Moral particularism*, pages 48–78. Oxford University Press.

Risse, M. (2002). Harsanyi's 'utilitarian theorem' and utilitarianism. *Noûs*, 36(4):550–577.

Roemer, J. E. (1996). *Theories of Distributive Justice*. Harvard University Press.

Roemer, J. E. (2008). Harsanyi's impartial observer is not a utilitarian. In Fleurbaey, M., Salles, M., and Weymark, J. A., editors, *Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls*, page 129–135. Cambridge University Press.

San, W. K. (ms). Aggregation and the structure of value. Retrieved from: personal communication.

Scanlon, T. (1998). *What We Owe to Each Other*. Belknap Press of Harvard University Press.

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica*, 57(3):571–587.

Segal, U. (1992). Additively separable representations on non-convex sets. *Journal of Economic Theory*, 56(1):89–99.

Sen, A. (1970). The impossibility of a Paretian liberal. *Journal of Political Economy*, 78(1):152–157.

Sen, A. (1973). *On Economic Inequality*. Oxford University Press.

Sen, A. (1976). Welfare inequalities and Rawlsian axiomatics. *Theory and Decision*, 7:243–262.

Sen, A. (1977). Non-linear social welfare functions. In Butts, R. and Hintikka, J., editors, *Foundational Problems in the Special Sciences. The University of Western Ontario Series in Philosophy of Science, Vol. 10*, pages 297–302. Springer.

Sen, A. (1979). Utilitarianism and welfarism. *Journal of Philosophy*, 76(9):463–489.

Sen, A. (1986). Social choice theory. In Arrow, K. J. and Intriligator, M. D., editors, *Handbook of Mathematical Economics*, pages 1073–1181. Elsevier.

Sen, A. (2017). *Collective Choice and Social Welfare: An Expanded Edition*. Harvard University Press.

Stefánsson, H. O. (2023). In defence of Pigou-Dalton for chances. *Utilitas*, 35(4):292–311.

Stefánsson, H. O. and Bradley, R. (2015). How valuable are chances? *Philosophy of Science*, 82(4):602–625.

Suppes, P. (1966). Some formal models of grading principles. *Synthese*, 16(3/4):284–306.

Suppes, P. and Winet, M. (1955). An axiomatization of utility based on the notion of utility differences. *Management Science*, 1(3/4):259–270.

Thoma, J. (2023). Taking risks on behalf of another. *Philosophy Compass*, 18(3):e12898.

Thomson, J. J. (1990). *The Realm of Rights*. Harvard University Press.

VanderWeele, T. J. (2017). On the promotion of human flourishing. *Proceedings of the National Academy of Sciences*, 114(31):8148–8156.

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

Voorhoeve, A. (2014). How should we aggregate competing claims? *Ethics*, 125(1):64–87.

Voorhoeve, A. (2021). Equality for prospective people: A novel statement and defence. *Utilitas*, 33(3):304–320.

Voorhoeve, A. and Fleurbaey, M. (2016). Priority or equality for possible people? *Ethics*, 126(4):929–954.

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

Wakker, P. (1991). Additive representations on rank-ordered sets. I. The algebraic approach. *Journal of Mathematical Psychology*, 35(4):501–531.

Wedgwood, R. (2022). The reasons aggregation theorem. *Oxford Studies in Normative Ethics*, 12:127–148.

Weirich, P. (2001). *Decision Space: Multidimensional Utility Analysis*. Cambridge University Press.

Weymark, J. (2005). Measurement theory and the foundations of utilitarianism. *Social Choice and Welfare*, 25(2/3):527–555.

Weymark, J. A. (1991). A reconsideration of the Harsanyi-Sen debate on utilitarianism. In Elster, J. and Roemer, J. E., editors, *Interpersonal Comparisons of Well-Being*, pages 255–320. Cambridge University Press.

Wilkinson, H. (2022). Can risk aversion survive the long run? *Philosophical Quarterly*, 73(2):625–647.

Williamson, T. (2007). *The Philosophy of Philosophy*. Wiley-Blackwell.

Zhao, M. (2021). Ignore risk; maximize expected moral value. *Noûs*, 57(1):144–161.

Zimmerman, M. J. (1983). Evaluatively incomplete states of affairs. *Philosophical Studies*, 43(2):211–224.