LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Doctoral Thesis

Essays in Decision Theory

Author: Zeev Avraham Goldschmidt

Supervisors: Prof. Richard Bradley and Prof. Christian List

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy
at the

Department of Philosophy, Logic and Scientific Method

July 21, 2024

Declaration of Authorship

I, Ze'ev Goldschmidt, certify that the thesis, titled Essays in Decision Theory, I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work, other than where I have clearly indicated that it is the work of others.

I confirm that Chapter 1 is coauthored with Christian List.

I declare that this thesis consists of 68,669 words.

Abstract

This thesis is composed of five essays tackling a range of issues in different areas of decision theory: moral uncertainty, the theory of reasons, and the semantics of taste predicates. In the first two chapters concerning moral uncertainty, I demonstrate how various versions of this problem are instances of broader phenomena addressed by general theories of rationality. Situating the problem of moral uncertainty in these broader contexts enables novel and illuminating applications of the general theories to it. In particular, the first two chapters demonstrate how the theories of belief binarization, and choice under indeterminacy – both theories devised independently of moral uncertainty – may be fruitfully applied to various versions of the morally uncertain agent.

The second two chapters explore two different aspects of the theory of reasons – their weight and the notion of acting for the right reasons. In chapter 3, I propose a method for measuring the weight of reasons that draws on the Theory of Measurement (Krantz et al. 1971). I derive a numerical representation of the weight of reasons from comparative judgments of outweighing relations, thus providing a rich structure rooted in intuitive judgments about the notion. In chapter 4, I apply the concept of acting for the right reasons to the case of assertion to argue that the norm of assertion is not epistemic. In doing so I demonstrate how concepts from the theory of reasons may have bearings on first-order issues like the contents of norms.

Finally, chapter 5 is about the semantics of taste predicates. I apply the theory of "Clouds of Contexts" developed with other purposes in mind (von Fintel and Gillies 2011) to allow the contextualist to account for linguistic intuitions arising from dialogues about taste. Such dialogues, on the proposed theory, function as explorations of individual and group taste-related attitudes.

Acknowledgements

In writing this thesis, I benefited greatly from the help, guidance, and encouragement of my supervisors Richard Bradley and Christian List to whom I am deeply thankful.

I would also like to thank my examiners, Caspar Hare and Campbell Brown, for giving this thesis such close attention, for their very helpful comments, and for a great philosophical conversation during my Viva.

I am very grateful to Andreas Achen, Campbell Brown, David Enoch, Liam Kofi Bright, Anna Mahtani, Ittay Nissan-Rozen, Mark Schroeder, and Andrew Schwartz, for very helpful comments and conversations about different parts of this thesis. I am also thankful to audiences at the LSE, LMU, and UC Berkeley for their thoughtful comments and questions.

I am extremely grateful to my mother, Penina Goldschmidt, for editing, believing, and nudging. Finally, I would like to thank my wife Rachel and my sons Yuval and Alon for being there for me in the most important ways.

Table of Contents

Introduction

This thesis is composed of five essays that concern different topics in decision theory broadly construed. The first two essays are about moral uncertainty, the second two are about the theory of reasons, and the final one is about the semantics of taste predicates.

Chapters 1 and 2 are about moral uncertainty. The problem of moral uncertainty concerns the relationship between an agent's beliefs regarding what they morally ought to do, and what they ought to do given those beliefs. The problem is typically set up by considering an agent who divides their credence between different moral hypotheses, or moral theories, that disagree about the moral status of some set of alternatives. How should such an agent choose among their alternative courses of action? The literature on moral uncertainty is concerned with this question, and more generally, with the norms governing the relationship between this kind of uncertainty and choice.

On a more abstract level, the problem of moral uncertainty is one of mapping epistemic states of a certain type to choice prescriptions, in a way that captures the appropriate practical response to those states. Thus framed, the problem of moral uncertainty is a decision-theoretic one, as decision theory is traditionally concerned with the norms constraining the relations between an agent's attitudes – like their beliefs and desires – and their choices or preferences.

The moral hypotheses, or moral theories the agent may be uncertain about, can vary considerably in their structure, and in the type of properties they employ to evaluate the relevant alternative courses of action. They could concern merely matters of permissibility, prohibition, and obligation, like statements of the form "it is permissible/prohibited/obligatory to choose x". Alternatively, the moral theories the agent considers possible may be *ordinal* – they may include pairwise "moral betterness" comparisons among alternatives – statements of the form "x is morally better than y" – and potentially fully rank alternatives in a "moral betterness" order. Finally, moral theories may be *cardinal* – they may allude to a notion of moral value that allows some alternatives to be morally *much* better, or merely *slightly* better, than others. Such theories are represented numerically on an interval scale, determining not merely the order of alternatives according to their moral value, but the ratios of the differences in moral value between them.

Morally uncertain epistemic states may thus be partitioned along these lines according to the structure of their objects – the type of moral hypotheses or theories about which the agent is uncertain. Indeed, much of the literature on moral uncertainty divides the problem in this way and considers uncertainty involving moral theories with different structures separately.[1] The first two chapters here will follow this practice, with the first tackling the ordinal case, and the second addressing the cardinal one. Both chapters identify the respective problems of moral uncertainty as instances of more general problems and draw on theories developed for those more general problems to yield novel insights about moral uncertainty.

In chapter 1, co-authored with Christian List, we consider the case of ordinal moral uncertainty. We introduce a formal framework for expressing the problem of ordinal moral uncertainty in which an agent divides their credence over propositions involving pairwise comparisons of moral betterness. We then demonstrate that the problem of

[1] For example, MacAskill et al. (2020) dedicate different sections of their book to these different types of moral uncertainty.

moral uncertainty resolution – the problem of mapping different credence functions to sets of propositions encoding choice prescriptions – is, formally speaking, an instance of the problem of *Belief Binarization* – the problem of mapping credence functions to sets of propositions the agent believes "outright" or "fully". This relationship allows the application of an impossibility theorem proved elsewhere (Dietrich and List 2018) with belief binarization in mind, to the context of ordinal moral uncertainty. The theorem states that any moral uncertainty resolution rule – any way of mapping morally uncertain credence functions to choice prescriptions – will violate at least one of four prima facie plausible conditions. We then consider the trade-offs involved in relaxing each of these four conditions, and taxonomize existing proposals in the literature according to the condition that they fail to satisfy.

Casting the problem of ordinal moral uncertainty in our framework facilitates the realization that it is a special case of a more general phenomenon of independent interest, namely, belief binarization. This diagnosis potentially allows applying a broad range of theories and insights developed for belief binarization to the problem of moral uncertainty. Our particular cross-application of the impossibility theorem allows mapping the space of possible solutions to the problem of moral uncertainty and sets the stage for analysing the theoretical trade-offs between different proposals in the literature.

Chapter 2 focuses on cardinal moral uncertainty, involving an agent dividing their credence between multiple cardinal moral theories, each evaluating the alternatives on an interval scale. The cardinal case introduces an extra complication, not present in the ordinal one, in the potential indeterminacy of Intertheoretic Value Comparisons (IVCs) – comparisons of value differences across the moral theories the agent considers possible. Much of the work on cardinal moral uncertainty agrees that at least in some cases IVCs are indeterminate and value differences are incomparable across moral theories. These types of cases are the focus of this chapter.

The chapter has two aims. First, to map the space of solutions to the problem of moral uncertainty with IVC indeterminacy, and to argue that the literature is unjustifiably skewed towards a certain type of solution. To do this, I draw on Williams' (2014) general framework for *choice under indeterminacy* – choice situations in which some indeterminacy in the normatively relevant facts render the normative status of some alternatives indeterminate. Applying this framework to the case of IVC indeterminacy demonstrates that a certain class of norms I term *Weak*, that is considered quite seriously in the debate of general indeterminacy, is largely overlooked as a possible solution in the context of moral uncertainty. Weak norms treat all ways of settling indeterminacy – and all prescriptions yielded by those ways – equally. The literature on moral uncertainty with IVC indeterminacy almost exclusively suggests norms that I term *Strong* – norms that favour certain ways of settling the indeterminate facts over other ways of doing so. I argue that this discrepancy in the way general indeterminacy and IVC indeterminacy are treated is unjustified, and Weak norms should be considered more seriously as solutions in the context of moral uncertainty.

The second aim of this chapter is to begin filling this lacuna by considering a particular Weak norm closely. In doing so, I prove that a natural Weak norm is equivalent to the norm prohibiting the choice of dominated alternatives – alternatives that are no better morally than some other alternative on all moral theories the agent considers possible, and are worse than it on some of them. This result gives rise to a novel argument for the weak norm that I consider, and it enables mapping the space of strong norms and the relations among them.

Like in chapter 1, situating the problem at hand in a more general framework developed independently of moral uncertainty yields novel insights about it – first about a type of solution that has been overlooked in the literature, and second an illuminating mapping of the space of solutions.

Both chapters demonstrate how different versions of the problem of moral uncertainty are instances of other, more general, problems studied in depth quite independently of moral uncertainty. In both chapters this realization – that the problem at hand is a special case of a more general one – allows importing theorizing, results, and insights developed elsewhere to the problem of moral uncertainty in ways that yield illuminating mappings of the space of possible solutions to the problem.

*

The second pair of chapters in this thesis concern different aspects of the theory of reasons. Normative reasons play a foundational role in all areas of normative theorizing. They appear in virtually any area where actions or attitudes are subject to norms – for example, practical reasons are relevant wherever norms of practical rationality apply, and epistemic reasons play roles in all areas governed by norms of epistemic rationality. While these are two central examples, the scope of reasons is much broader – there are reasons for many other attitudes like fear, excitement, concern, and many more. Perhaps the most important feature of reasons is their relation to what one ought to do – whether one ought to ϕ[2] is closely related to certain relations holding between the reasons for and the reasons against ϕ-ing. But reasons play a broad range of other roles in normative theorizing, and rich theories of reasons have been developed that demonstrate how rich

[2] I use ϕ as a placeholder for an arbitrary action or attitude.

and pervasive this concept is.[3] The two chapters about reasons address two different features of reasons developed in these theories – their weight and their being objects of motivation, or more specifically, the notion of acting for the right reasons.

Chapter 3 addresses the notion of the weight of reasons and proposes a way of representing it numerically. The weights of reasons feature saliently in the phenomenology of deliberation and play an important role in normative theorizing. When deliberating whether to ϕ – whether to act in a certain way or adopt a certain attitude – it is typical to weigh the reasons for and against ϕ-ing. Indeed, such weighing and comparing is a central part of figuring out what one ought to do. Weight also figures prominently in normative theorizing, in explaining the relation between reasons and other normative properties. Most centrally, whether one ought to ϕ appears to be closely linked to, and perhaps explained by, whether the reasons for ϕ-ing outweigh the reasons against doing so.

Our intuitive notion of the weight of reasons appears to have a considerably rich structure as it makes sense not merely to talk about some reasons being weightier than others, but also of some reasons being *much weightier* or *slightly weightier* than others. However, it is not obvious whether the notion may be represented numerically, and if it may, how precisely it ought to be represented. In particular, assigning numbers to sets of reasons to represent their weight may initially seem to impose too rich a structure on the notion of weight, and to raise a host of vexing questions regarding the meaning of such assignments and the type of facts that make them correct.

[3] I am mostly influenced by Schroeder (2007,2021a) here, but see also Parfit (1984) and Scanlon (2014).

In this chapter, I propose a method for measuring the weight of reasons in a way that alleviates these worries. I propose deriving a numerical representation of weight from numberless – and, I argue, intuitive – comparative judgments concerning weight and outweighing relations. I construct a formal model for expressing these relations and apply a theorem from the Theory of Measurement (Krantz et al. 1971) to derive a numerical representation of weight. The theorem demonstrates that if outweighing relations among sets of reasons satisfy certain prima facie plausible conditions, then weight may be represented numerically. The derivation allows understanding the numerical side of the representation in terms of the underlying outweighing relations from which they are derived. I discuss the conditions ("axioms") required by the theorem and consider the implications of weakening some of them. Finally, I discuss the possible shortcomings of the derived representation especially in relation to its partial ability to account for reason *accrual* – the relationship between the weights of sets of reasons and the weights of the individual reasons of which they are composed.

The chapter begins to close a gap in the literature on the weight of reasons, where it is often assumed that reasons may be represented numerically, but no story about the origins or plausibility of such a representation is told (Brown 2014, Sher 2019, Nair 2021). The chapter may also serve to illuminate the relationship between decision theory and the theory of reasons, especially the parts about practical and epistemic reasons. While the Theory of Measurement plays a frequent and central role in decision theory – a field in which representation theorems for measuring both belief and desire abound – it is much less centrally featured in the study of reasons.[4] Applying the theory of measurement to the weight of reasons has the potential of illuminating the relationship between the weights of, e.g., practical reasons, and a rational agent's utility function, and

[4] With the exception of Dietrich and List (2013).

the weights of epistemic reasons and a rational agent's credence function. While I do not pursue this direction here, the application of methods of measurement to the field of reasons may help to do so in future work.

Chapter 4 leverages the theory of reasons, and in particular, the notion of *ϕ-ing* for the right reasons, to weigh in on a debate about the norm of assertion. In particular, I argue that applying this notion of ϕ-ing for the right reasons to assertion reveals that the norm of assertion cannot be epistemic – it cannot be of the form "assert p only if you stand in epistemic relation R to p". The argument relies on Schroeder's notion of *ϕ-ing well* – being motivated to ϕ (rightly) by enough of the reasons that render ϕ-ing right – and is composed of two parts. First, I demonstrate that if the norm of assertion is epistemic then asserting well requires one to be motivated to assert by facts about one's mental states. Second, I argue that asserting well cannot require this motivation, and therefore the norm of assertion cannot be epistemic. I then offer several debunking arguments explaining the misleading allure of epistemic norms for assertion. Finally, I offer a preliminary positive account for the norm of assertion.

The chapter demonstrates how concepts from the theory of reasons – like ϕ-ing well – may have bearings on first-order issues like the content of the norm of assertion. The method employed in this chapter, of evaluating a norm by examining the *well standard* it gives rise to, may generalize to other norms in other contexts.

The theory of reasons and decision theory partially overlap. In particular, both decision-theoretic norms and norms of practical reasoning constrain the same objects, namely, agents' choices or preferences. Orthodox decision theory requires that the agent choose in a way that maximizes expected utility, and norms of practical reasoning would require agents to make choices that are supported by, or stand in other normative relations to, certain sets of reasons. There is much to explore about how these disparate yet overlapping disciplines relate to each other, and whether their prescriptions are always consistent. Chapters 3 and 4 may be seen – in addition to achieving their explicit aims – as laying some of the groundwork for such an exploration. Proposing a numerical representation of the weight of reasons couples the propositional notion of a reason, with a cardinal notion of weight. Since utility and probability – two fundamental decision-theoretic entities – are cardinal, the "cardinalisation" of reasons brings the two fields closer, facilitating the exploration of the ways they may relate to each other (without resorting to reducing reasons to probability or utility). Second, the application of the notion of ϕ-ing well to assertion in chapter 4, demonstrates that this notion may have important implications for normative domains not typically articulated in terms of reasons. This paves the way for using this framework to explore how decision-theoretic norms may relate to motivational constraints formulated in terms of reasons in general, and how such norms may be evaluated by examining the well standards they give rise to in particular.

*

Finally, the last chapter of this thesis is about the semantics of taste predicates, predicates like "tasty" and "fun". Contextualist theories for taste predicates take them to be perspective-relative and require the context to fix the relevant perspective for their evaluation. On this picture, saying that something is tasty is saying that something is tasty *to* some contextually-determined individual or group. However, dialogues involving disagreement about taste generate problems for such theories as there seems to be no way of fixing the perspectives in question while respecting linguistic intuitions. In particular, the contextualist appears unable to account for the intuition that such dialogues involve faultless disagreement – genuine disagreement in which no party is at fault. This chapter proposes a solution to this problem by applying von Fintel and Gillies' (2011) theory of "Clouds of Contexts" to the case of taste predicates. Their theory was originally designed to solve a problem for the contextualist in relation to epistemic modals but, I argue, is also able to solve their problems relating to taste predicates, due to the similarity between these two linguistic phenomena.

Applying the Cloud of Contexts theory to the case of taste predicates allows assertions involving such predicates to carry, or "bring into play", multiple perspectives – not merely the speaker's but also, possibly, the perspectives of other individuals and groups. This introduces a degree of flexibility to the meaning of taste predicates that enables accounting for dialogues involving agreement and disagreement about matters of taste.

On the contextual picture defended in this chapter, taste predicates express individual and/or group attitudes, and conversations involving such predicates function as instruments for exploring and debating these attitudes. An utterance involving a taste predicate may evoke a broad range of perspectives, introducing to the conversational agenda the perspectives of the speaker, the listeners, the possible groups composed of them, the average member of some group of reference, and more objective perspectives like those of some contextually determined experts. The ensuing conversation then allows interlocutors to jointly discover and debate the attitudes of each of the relevant perspectives. On a more abstract level, the story told here about taste predicates is one of the exploration of individual and group attitudes and the ways they relate. In this sense, the study of taste predicates has affinities to decision theory, a field focused on accounting for other individual and group attitudes, the relations among them, and the normative constraints they are subject to.

However, the links between taste predicates and decision theory run deeper. Taste predicates appear to have a lot in common with desire and preference, two attitudes at the core of decision theory. Both taste predicates like "tasty" and the predicates "preferable" and "desirable" are perspective-relative – something may be preferable, desirable or tasty to me, but not to you. These predicates also have more objective readings – in some contexts, saying that something is preferable or desirable, conveys an objective sense of preferability or desirability that is independent of the speaker's perspective – think of a policy maker stating that some intervention is preferable, or some outcome is desirable. Similarly, in certain contexts, taste predicates are best understood in a more objective way – consider for example, a food critic deeming some dish extremely tasty. In addition, both types of predicates arguably express attitudes that may be held by groups as well as individuals, and the relations between the individual and group attitudes raise similar theoretical questions. Finally, both classes of predicates appear in the same linguistic constructions. They may be found in "bare" form: "x is tasty (preferable, desirable)"; with a "to" or "in my opinion" argument: "x is tasty (preferable, desirable) to me/ in my opinion"; embedded in propositional attitude verbs like "find": "I find x tasty (preferable, desirable)";[5] and finally, in the comparative form, like: "x is tastier than (preferable to, more desirable than) y" (the other constructions above can also be formulated in comparative form). While I do not pursue these parallels in the chapter, I believe that the progress it makes in relation to taste predicates is applicable, at least in part, to the understanding of the attitudes of preference and desire, attitudes at the core of decision theory.

[5] See Crespo and Fernández (2011).

*

In sum, this thesis includes analyses of different areas in decision theory broadly construed: moral uncertainty in its various forms, the theory of reasons and its implications, and the meaning of taste predicates. However, on a more general level, the thesis is a project in philosophical consolidation. It demonstrates, in different ways in each of its chapters, how the realization that a certain philosophical problem is an instance of another, more general one, allows applying theories across contexts in a fruitful way. This philosophical manoeuvre allows shedding new light on the problems under consideration and importing insights from one philosophical area to another. It enables making use of a rich body of theoretical work developed independently of the problem of interest, to yield a new understanding of it and of its potential solutions.

This philosophical method is employed throughout the thesis. In chapter 1, demonstrating that the problem of ordinal moral uncertainty is formally an instance of belief binarization enables drawing insights and results from the latter to the former. In chapter 2, applying the theory of choice under indeterminacy to moral uncertainty with indeterminate IVCs, draws attention to a type of solution that is prominent in the general case, but overlooked in the moral uncertainty literature. Drawing on the theory of measurement in chapter 3, allows making progress on the numerical representation of the weight of reasons. In chapter 4, applying the notion of acting for the right reasons to assertion, yields new lessons about the content of the norm governing assertion. Finally, in chapter 5, closely examining the well-established similarities between taste predicates and epistemic modals[6] enables the application of the Cloud of Contexts theory, originally developed with the latter in mind, to the former.

This thesis may thus be seen as an argument for the fruitfulness of this kind of cross-pollination, and for the utility of applying theories across contexts in the manner pursued here. It may also be read as an invitation to further pursue the connections revealed and made fruitful here, to yield new insights about the considered topics.

[6] See Stephenson (2007) and Schaffer (2011).

Chapter 1

Moral Uncertainty Resolution as Belief Binarization: An Impossibility Result

Coauthored with Christian List

Moral uncertainty occurs when we are uncertain about moral matters. We may divide our credence among competing moral views and only assign credences strictly between 0 and 1 to propositions such as "action x is morally better than action y". What action-guiding judgments should we rely on in such cases? In this chapter, we approach this question from a novel angle. We suggest that the problem of "moral uncertainty resolution" can be viewed as a special case of the problem of "belief binarization": the problem of arriving at all-out ("accept/reject", "yes/no") judgments on some propositions based on our assignment of credences to them. Looking at moral uncertainty through this lens yields a diagnostically useful impossibility result, which shows that there is no moral uncertainty resolution rule that satisfies four initially plausible conditions. The theorem allows us to map out the space of possible solutions in a congenial way, demonstrating how any solution to the problem of moral uncertainty comes with some theoretical cost.

1. Introduction

Some agents are sometimes uncertain about moral matters. For example, they may be uncertain about whether animal suffering is as bad as human suffering, whether banning unvaccinated people from restaurants is morally permissible, how much risk one may justifiably impose on future generations to achieve some present benefit, or how important it is to protect nature for its own sake. Those agents will then be uncertain about the moral status – the choice-worthiness, value, or permissibility – of some of the available courses of action. We may think of such agents as dividing their credence among competing first-order moral viewpoints (or moral theories). When it comes to choosing between different actions, what action-guiding judgments should such agents rely on?

In this chapter, we approach this question from a novel angle. We suggest that the problem of moral uncertainty resolution can be fruitfully viewed as a special case of the problem of belief binarization: the problem of arriving at all-out judgments on some propositions – of an "accept"/"reject" or "yes"/"no" form – on the basis of an assignment of credences to these propositions (see, e.g., Leitgeb 2014 and Dietrich and List 2018). For example, if our credences in p, q, and $p\ and\ q$ are 0.9, 0.8, and 0.72, respectively, the belief-binarization task is to determine what all-out beliefs we should have: which, if any, of these three propositions – p, q, and $p\ and\ q$ – should we accept, especially for action guidance? We show that looking at moral uncertainty through the lens of belief binarization yields a diagnostically useful impossibility result, namely: if we require any solution to our problem – a "moral uncertainty resolution rule" – to satisfy four prima facie plausible conditions, called "universality", "ordering", "certainty preservation", and "propositionwise independence", then there exists no such rule. The upshot is that, to find a moral uncertainty resolution rule, we must relax at least one of those conditions. This observation enables us to taxonomize different solutions to our problem according to the conditions they relax.

Furthermore, since the theory of belief binarization on which we build (Dietrich and List 2018) is itself a novel application of the theory of judgment aggregation in social choice theory, our analysis also sheds new light on the connections between moral uncertainty resolution and social choice. Some of those connections have previously been explored in the literature, especially by MacAskill (2016), but the connection is usually made with traditional voting theory in the tradition of Nicolas de Condorcet and Kenneth Arrow, not judgment-aggregation theory. Neither belief binarization nor the precise theorem presented below feature in the earlier analyses. Although our main impossibility result follows directly from an existing impossibility result on belief binarization (Dietrich and List, 2018), its application to moral uncertainty is new and should be of interest in its own right, as should be the identification of the parallel between moral uncertainty resolution and belief binarization.

In Section 2, we narrow down our focus and introduce the version of the problem of moral uncertainty resolution that we will be mostly concerned with: ordinal moral uncertainty resolution. In Section 3, we present the formal framework. In Section 4, we explain why so-called "threshold rules", which may initially seem obvious solutions, don't generally work. In Sections 5 and 6, we state our impossibility result and discuss possible escape routes. In Section 7, we explain how our analysis can be generalized beyond the ordinal version of the problem of moral uncertainty.

2. Narrowing Down Our Focus

For our analysis, we will mostly focus on the ordinal version of the problem of moral uncertainty resolution. That is, we will assume that each of the competing moral viewpoints ("moral theories") about which an agent is uncertain takes the form of a ranking of the relevant objects of moral assessment (actions, states, consequences etc., for short alternatives) in an order of choice-worthiness, goodness, or desirability. In the first place, the agent's state of uncertainty will be expressed by a credence distribution over the set of possible such orders. For instance, the agent might assign a credence of $1/3$ to the order which places alternative x above alternative y above alternative z, $1/3$ to the order which places y above z above x, and $1/3$ to the order which places z above x above

y. Could the agent then arrive at an all-things-considered order of the alternatives – an appropriateness order – on the basis of which they should choose, given their moral uncertainty?[1] For generality, we will also consider the case in which the agent doesn't assign credences to complete rankings of the alternatives but just to binary comparisons between alternatives. For instance, the agent might assign a credence of $2/3$ to the proposition that x is morally better than y, a credence of $2/3$ to the proposition that y is morally better than z, and so on; and the agent might then want to figure out whether to choose x over y, y over z, x over z, and so on.

Our aim is to investigate whether there exists a reasonable *moral uncertainty resolution rule*, a decision rule which takes the agent's credal state as input and produces as output an appropriateness order of the alternatives, which can guide the agent's choices. As will become evident, this is formally an instance of a *belief-binarization rule*, a function that takes credences on some set of propositions as input and produces determinate "accept"/"reject" verdicts on these propositions as output (Dietrich and List, 2018). In the context of belief binarization, binary "acceptance" verdicts are usually interpreted as full beliefs in the given propositions. In the current context, they are best interpreted as appropriateness verdicts for the purposes of action guidance. Despite this interpretative difference, the formal similarity allows applying the work on belief binarization to the current context.

The present *ordinal* version of the problem of moral uncertainty is distinct from the *cardinal* version, which has received much more attention in the philosophical debate. In that cardinal version, the competing moral theories among which an agent distributes

[1] We adopt the term "appropriateness" from MacAskill (2016). It is controversial whether the normativity involved in what one ought to do given first-order moral uncertainty is that of rationality (Sepielli 2009) or morality (Rosenthal 2021). We remain neutral on this issue, as our framework is open to both interpretations.

their credence are each represented, not merely by an *order* of the alternatives, but by a *cardinal value function*, which assigns a real number to each alternative, expressing its value according to the given theory. Under the assumption that these values are measurable on an interval scale *and* comparable across different theories, some philosophers have argued that the appropriate response to a state of moral uncertainty is to rank the alternatives in an order of their *expected value*, where the expectation is calculated on the basis of the agent's credence distribution over the competing theories.[2] This effectively extends the standard decision-theoretic recommendation of expected-utility maximization from the familiar case of empirical uncertainty to the new case of moral uncertainty. The biggest difficulty with this proposal, however, is that it is unclear whether many moral theories outside the relatively narrow class of utilitarian theories and perhaps some other consequentialist theories admit the required cardinal representation with values measurable on an interval scale. It is even less clear whether comparisons of such values across distinct theories make sense. Could one really say, for instance, that a switch from not lying to lying according to a Kantian theory is exactly equivalent to a particular numerical utility loss – say, of 100 utiles – according to a Benthamite utilitarian theory? We cannot settle these questions here, but it seems to us that the burden of proof is on those who think that the cardinal case of moral uncertainty should be viewed as the default case. After all, the cardinal case is both informationally and interpretationally much more demanding than the ordinal one.[3]

[2] For arguments in favour of this view, see Lockhart (2000b), Sepielli (2010), MacAskill and Ord (2020), Riedener (2020), and Carr (2020). For arguments against it, see Nissan-Rozen (2015b) and Hedden (2016b). For a distinction between different formulations of this view when both moral and empirical uncertainty are involved see Dietrich and Jabarian (2022).

[3] On some views, the required response to moral uncertainty is independent of whether the theories under consideration are ordinal, cardinal, or of some other structure. For example, the “My Favourite Theory” (Gustafsson and Torpman 2014) and “My Favourite Option” decision rules are sensitive only to the ordinal content of the theories under consideration regardless of their additional structure (see section 6.4). The ordinal/cardinal distinction is also inconsequential for the *Normative Externalist* (Weatherson, 2014, 2019a)

In any event, since at least some moral theories have an ordinal rather than cardinal format, it is interesting to investigate the problem of ordinal moral uncertainty resolution in its own right. Indeed, the class of ordinal moral theories is quite broad. It includes many deontological or virtue theories that only rank options according to their alignment with some set of deontic constraints (e.g., obligations and permissions) or virtues, and do not make use of richer metrics such as cardinal differences or degrees of choice-worthiness. More generally, any normative theory that yields a choice function satisfying certain consistency constraints, where that choice function encodes which alternatives from any given "menu" are permissible, admits an ordinal representation and can thus be viewed as an ordinal normative theory for the purposes of our analysis.[4] In Section 7, we explain how our analysis can be extended to theories whose format differs from the ordinal one, including cardinal theories and theories expressible in a logic with deontic operators or evaluative predicates.

3. Our Framework

Let A be some non-empty and finite set of objects of moral assessment: these may be courses of action, alternative policies, or alternative states of affairs. We simply call them *alternatives*. Let Ω denote the set of all logically possible complete orders $\succcurlyeq$ (transitive and complete binary relations) on A, where, for any alternatives $x, y \in A$, $x \succcurlyeq y$ is interpreted to mean that x is at least as morally good or valuable, or at least as choice-worthy, as y.[5]

– the proponent of the view that what one ought to do given moral uncertainty is determined by the first-order moral facts, and independently of one's beliefs about such facts. See Tarsney (2021) for an extensive discussion of theory structure and its consequences for moral uncertainty.

[4] Relevant decision-theoretic representation results (albeit without an explicit moral application) are surveyed in Bossert and Suzumura (2010). An application of decision-theoretic representation theorems to the moral domain can also be found in C. Brown (2011).

[5] Transitivity means that, for all $x, y, z \in A$, if $x \succcurlyeq y$ and $y \succcurlyeq z$, then $x \succcurlyeq z$. Completeness means that, for all $x, y \in A$, $x \succcurlyeq y$ or $y \succcurlyeq x$ (or both).

For instance, when there are three alternatives in A, there are 13 logically possible complete orders on them, as shown in Table 1.1, and Ω will then consist of those 13 orders.[6] We write $x \succ y$ as a shorthand for $x \succcurlyeq y$ and not $y \succcurlyeq x$, and $x \approx y$ as a shorthand for $x \succcurlyeq y$ and $y \succcurlyeq x$.

TABLE 1.1: Logically possible complete orders over three alternatives ("moral worlds")

1	$x \succ y \succ z$
2	$x \succ z \succ y$
3	$y \succ x \succ z$
4	$y \succ z \succ x$
5	$z \succ x \succ y$
6	$z \succ y \succ x$
7	$x \succ y \approx z$
8	$y \succ x \approx z$
9	$z \succ x \approx y$
10	$x \approx y \succ z$
11	$x \approx z \succ y$
12	$y \approx z \succ x$
13	$x \approx y \approx z$

We may think of each element of Ω as representing a possible moral theory, or as encoding a complete set of moral facts about the relative goodness or choice-worthiness of the alternatives in A. The set Ω can thus be interpreted as the set of all possible "moral worlds", where a "moral world", for present purposes, is a complete specification of the ordinal moral facts about the alternatives in A.

We can further think of any subset of Ω as encoding a moral proposition. For instance, the subset consisting of all the orders in which we have $x \succcurlyeq y$ corresponds to

[6] Of these, six involve no ties between any distinct alternatives, while the rest involve ties between at least two or even all three alternatives.

the proposition “x is at least as good/choice-worthy as y”. The subset consisting of all the orders in which we have $x \succcurlyeq y$ and not $y \succcurlyeq x$ corresponds to the proposition “x is strictly better (more choice-worthy) than y” ($x \succ y$).

Logical operations on moral propositions are easily definable. To form the *conjunction* (logical “and”) of any two moral propositions, represented by the subsets $p, q \subseteq \Omega$, we take the intersection $p \cap q$. To form their *disjunction* (logical “or”), we take the union $p \cup q$. To form the negation of any moral proposition p, we take its set-theoretic complement $\neg p := \Omega \backslash p$. The set of all moral propositions – formally the set of all subsets of Ω – thus constitutes a Boolean algebra; call it $\mathcal{P}(\Omega)$.[7] Within this algebra, a set S of propositions is logically consistent if the intersection of its members is non-empty (i.e., $\cap_{p \in S}\, p \neq \emptyset$); and a set S of propositions entails another proposition q if the intersection of the propositions in S is a subset of q (i.e., $\cap_{p \in S}\, p \subseteq q$).

We can now represent an agent's state of moral uncertainty in terms of a credence function on some set X of moral propositions of interest, where X is some non-empty subset of $\mathcal{P}(\Omega)$ that is closed under negation. The set X will be called the *agenda*; we can think of it as the set of propositions to be adjudicated. Formally, a *credence function* on X is a real-valued function Cr (from X into [0,1]) which assigns to each moral proposition $p \in X$ a number $Cr(p)$ between 0 and 1, understood as the agent's credence in p, subject to the constraint of probabilistic coherence.[8]

Our default case is $X = \mathcal{P}(\Omega)$, i.e., the agent assigns credences to all moral propositions in the given algebra. But we will also consider the case in which X is

[7] It is closed under negation, conjunction, and disjunction, as just defined.

[8] Probabilistic coherence requires that Cr be extendable to a function Cr^* on $\mathcal{P}(\Omega)$ with the properties that (i) $Cr^*(\Omega) = 1$ and (ii) for any $p, q \in \mathcal{P}(\Omega)$ with $p \cap q = \emptyset$, $Cr^*(p \cup q) = Cr^*(p) + Cr^*(q)$.

restricted to moral propositions of the form $x \succcurlyeq y$ and their negations, i.e., X is defined as the set:

$$\mathcal{P}_{binary} := \{\, x \succcurlyeq y, \neg(x \succcurlyeq y) : x, y \in A\}$$

If $X = \mathcal{P}_{binary}$, then the agent assigns credences only to betterness comparisons between pairs of alternatives (and to the negations of such comparisons), but not to other moral propositions, such as logical combinations of pairwise betterness comparisons.[9]

We are now in a position to define a moral uncertainty resolution rule. This is a function f that assigns to each admissible credence function Cr on X a corresponding "appropriateness judgment set" J consisting of all those moral propositions (from amongst the ones in X) that are "accepted" for action-guiding purposes, given the agent's moral uncertainty. Formally, f is an instance of a belief binarization rule, as defined in Dietrich and List (2018), a function that converts credences on some propositions into "all-out" beliefs or acceptances. Of course, an appropriateness judgment need not be interpreted as an all-out belief. Rather, they should be understood as representations of appropriateness facts – facts about what one ought to choose given one's moral uncertainty, or more generally, facts about the metanormative status of the alternatives.

The binary appropriateness judgment set J is called:

- *consistent* if it is a logically consistent set;
- *complete* (in X) if it contains a member of every proposition-negation pair $p, \neg p \in X$;

[9] In MacAskill's (2016) framework, credences are defined only over moral theories, i.e., the elements of Ω (more precisely, its singleton subsets), but not over moral propositions (subsets of Ω) more generally. In Tarsney's (2019) framework, credences are defined over elements of Ω and over $\mathcal{P}_{binary}$. Our framework is therefore more general than both as it permits credences to be defined, in principle, over any non-empty, negation-closed subset of $\mathcal{P}(\Omega)$.

- *deductively closed* (in X) if it contains every proposition $q \in X$ that it entails (i.e., if J entails q, then $q \in J$).

Consistency and completeness jointly imply deductive closedness. It is worth noting that, regardless of whether X is $\mathcal{P}(\Omega)$ or $\mathcal{P}_{binary}$, a consistent and complete (and thereby deductively closed) judgment set $J \subseteq X$ encodes a unique complete order $\succcurlyeq$ on A, which is defined as follows: for any $x, y \in A$,

$$x \succcurlyeq y \text{ if and only if } (x \succcurlyeq y) \in J.$$

The order $\succcurlyeq$ thus defined is also the unique order in Ω that lies in the intersection of all the propositions in J. By contrast, if the judgment set J is consistent and deductively closed, but not necessarily complete, then the binary relation $\succcurlyeq$ thus defined is a partial order on A (a reflexive and transitive but not necessarily complete binary relation). Finally, a merely consistent judgment set J – not necessarily complete or deductively closed – encodes what is technically known as a Suzumura consistent binary relation on A: one that is free from any pairwise comparison cycles in which at least one comparison is strict. An example of such a cycle, which would violate Suzumura consistency, is $x \succcurlyeq y$, $y \succcurlyeq z$, and $z \succ x$.

One reason why arriving at a complete order – or at least a cycle-free binary relation – on the set A of alternatives may be important is that the agent may need it for action guidance. Given a choice among a menu of alternatives, the agent may always wish to choose a maximally appropriate alternative. It is well-known in the literature on rational choice (e.g., Arrow, 1959, Bossert and Suzumura, 2010) that if the resulting choice function is required to satisfy certain standard consistency constraints (such as "contraction consistency" or the "weak axiom of revealed preference"), this requires – and indeed is logically equivalent to – having an underlying binary relation over the alternatives that is sufficiently well-behaved, for instance a complete order or at least a cycle-free relation.

4. Why Threshold Rules Won't Work

One intuitively natural class of moral uncertainty resolution rules corresponds to the class of threshold rules for belief binarization. A threshold rule specifies a particular credal threshold t between 0 and 1 and deems any proposition $p \in X$ acceptable if and only if $Cr(p)$ exceeds that threshold t. Here, "exceeds t" can be defined either "weakly" as $Cr(p) \geq t$ or "strictly" as $Cr(p) > t$. Formally, a threshold rule, with threshold t, is a function f from the set of all credence functions on X to the set of subsets of X that assigns to each credence function Cr the judgment set:

$$J = \{\, p \in X \colon Cr(p) \text{ exceeds } t\},$$

with a given interpretation of "exceeds" (as "≥" or ">"). So, if $X = \mathcal{P}_{binary}$, for instance, the pairwise comparisons $x \succcurlyeq y$ that are accepted for action guidance are all and only those the agent gives a credence that exceeds the relevant threshold t.

Threshold rules are appealing because they appear to preserve some degree of alignment between the agent's moral beliefs and the actions licensed by those beliefs. For example, a threshold rule with a strict threshold of 0.5 will ensure that, for any two alternatives x and y, x is deemed at least as appropriate as y if and only if $Cr(x \succcurlyeq y) > 0.5$, i.e., if and only if the agent finds it more likely than not that x is at least as choice-worthy as y. An agent abiding by such a threshold rule would never be in the uncomfortable position of having their choices divorced from their credences regarding choice-worthiness in the sense of choosing y over x without assigning a sufficiently high credence to y's being at least as choice-worthy as x.

The trouble with threshold rules, however, is that they do not generally produce a consistent and complete appropriateness judgment set, which would encode an ordering of the alternatives, nor even a judgment set that is consistent and deductively closed but possibly incomplete, which would encode a partial ordering of the alternatives. First of all, for any threshold above 0.5, it can easily happen that neither $x \succcurlyeq y$ nor $y \succcurlyeq x$ receives a credence assignment above the relevant threshold, so that the agent would not be able to make any overall betterness comparison between x and y – an instance of incompleteness of the appropriateness judgment set. Furthermore, it can also happen that the agent assigns a credence above the threshold to $x \succcurlyeq y$ and also to $y \succcurlyeq z$, but not to $x \succcurlyeq z$, so that the agent's overall comparisons between the alternatives would be intransitive – a violation of deductive closedness of the judgment set. This is, in effect, the moral uncertainty analogue of the "lottery paradox" for belief binarization.

For a simple example, suppose the credal threshold is $2/3$ (weakly defined as "$\geq 2/3$"), and suppose the agent assigns an equal credence of $1/3$ to each of the following three betterness orders of the alternatives:

$$x \succ y \succ z,$$

$$y \succ z \succ x$$

$$z \succ x \succ y,$$

Then $x \succcurlyeq y$ and $y \succcurlyeq z$ would each be assigned a credence of $2/3$, thereby meeting the threshold for acceptability, while $x \succcurlyeq z$ would not – a violation of transitivity.

What's even worse, in this example, $z \succcurlyeq x$ would also be assigned a credence of $2/3$, so that the resulting overall comparisons would be cyclical – in fact, with strict binary comparisons, i.e., $x \succ y$, $y \succ z$, and $z \succ x$ – a violation of (Suzumura) consistency. Readers familiar with Condorcet's classic voting paradox in social choice theory will recognize that the three orders in the present example are identical to those in Condorcet's example of cyclical majority preferences. Thus, we have not just a formal analogue of the lottery paradox, but also of Condorcet's paradox.[10] The lesson is that we cannot generally rely on threshold rules for resolving ordinal moral uncertainty.[11]

5. An Impossibility Result

Since threshold rules are not generally satisfactory, it is reasonable to ask whether there could be other moral uncertainty resolution rules that work better. In what follows, we show that if we impose four at first sight plausible conditions on a moral uncertainty resolution rule f, then the answer is negative. We will then need to discuss which, if any, of these conditions might be relaxed. Here are the four conditions:

> **Universality**. The domain of admissible inputs to f is the set of all possible credence functions on X.
>
> **Ordering.** The output of f is always a consistent and complete appropriateness judgment set $J \subseteq X$, or equivalently a complete appropriateness order of the alternatives in A.

[10] A previous discussion of the moral uncertainty analogue of the Condorcet paradox may be found in Nissan-Rozen (2012), who argues for the use of lotteries in such cases.

[11] Tarsney (2018) argues for a threshold rule for the morally uncertain deontologist, who is uncertain whether alternatives violate deontological constraints. The proposed threshold rule requires agents to choose an option x only if their credence in the proposition "x does not violate any deontological constraint" is above some threshold t. The rule indeed runs into an analogue of the "lottery paradox" which Tarsney terms (following Jackson and Smith 2006) "failure of *ought agglomeration*": in some cases, the agent's credence in the propositions "x does not violate any deontological constraint" and "y does not violate any deontological constraint" both exceed t, while the agent's credence in their conjunction does not. Tarsney escapes this problem and preserves *ought agglomeration* by limiting the threshold rule to simple options. See his "Minimalism". See also Aboodi et al. (2008).

Certainty preservation. For any credence function Cr in the domain of f, if Cr is uncertainty-free – i.e., it assigns credences 0 or 1 to all propositions in X – then $J = f(Cr)$ consists of precisely those propositions $p \in X$ with $Cr(p) = 1$.

Propositionwise independence. For any credence functions Cr and Cr' in the domain of f and any $p \in X$, if $Cr(p) = Cr'(p)$ then $p \in J \Leftrightarrow p \in J'$, where $J = f(Cr)$ and $J' = f(Cr')$.

Universality and ordering should be fairly self-explanatory requirements if the aim is to find a moral uncertainty resolution rule that is generally applicable and that resolves the given uncertainty in a decisive way. Certainty preservation should also be self-explanatory: in the special case of no uncertainty (where the credences take only the extremal values 0 or 1), the implied judgments can be read off from the credences in the obvious way. Propositionwise independence, finally, says that the action-guiding judgment on any proposition p should depend only on the credence in p, not on the credences in other propositions. In the special case in which $X = \mathcal{P}_{binary}$, this reduces to the requirement that the agent's overall judgment on any "pairwise comparison proposition" of the form $x \succcurlyeq y$ or $\neg(x \succcurlyeq y)$ depends only on the agent's credence in that specific pairwise comparison proposition itself, not on the agent's credences in distinct pairwise comparison propositions, for instance ones involving other alternatives. We will later discuss the possibility of relaxing each of these conditions.

The following result is an immediate application of the main theorem in Dietrich and List (2018).

Theorem 1.1: If $|A| \geq 3$, the four stated conditions are inconsistent (regardless of whether X is $\mathcal{P}(\Omega)$ or $\mathcal{P}_{binary}$), i.e., there exists no moral uncertainty resolution rule satisfying universality, ordering, certainty preservation, and propositionwise independence.

It is worth noting that this theorem has a precursor in social choice theory and specifically in the theory of judgment aggregation. Suppose we reinterpret a credence function Cr on the set X as a function that assigns to each proposition $p \in X$ the proportion of individuals within a given electorate or committee who accept proposition p (e.g., who make a "yes" or "true" judgment on p). Then the function Cr represents a particular anonymous profile of judgments or opinions on the given propositions within the electorate or committee in question. (The profile is called *anonymous* because it does not specify which individuals or voters support which propositions; it only specifies proportions of support for them.) For instance, it might encode information of the form: $2/3$ of the voters or committee members accept proposition p, $1/3$ accept proposition q, no-one accepts proposition r, and so on. The uncertainty resolution rule (belief-binarization rule) that assigns to each function Cr a judgment set $J \subseteq X$ can then be reinterpreted as an *anonymous judgment aggregation rule*, which takes proportions of support on the various propositions as input and produces a set of collectively accepted propositions as output. Given this interpretation, the theorem says that, under the stated assumptions about X, there exists no such aggregation rule satisfying the social-choice-theoretic variants of the four stated conditions (Dietrich and List 2007, Dokow and Holzman 2010; for precursors, see List and Pettit 2002 and Nehring and Puppe 2002). In this social-choice-theoretic interpretation, universality requires that any anonymous profile of individual judgments be acceptable as input; ordering requires that collective judgments be consistent and complete; certainty preservation becomes unanimity preservation; and propositionwise independence requires that the collective judgment on each proposition should depend on individual judgments on that proposition alone. The last condition is a judgment-aggregation variant of Kenneth Arrow's much-discussed (albeit controversial) condition of the "independence of irrelevant alternatives". The impossibility of simultaneously satisfying these four conditions, in turn, is a judgment-aggregation analogue of Arrow's (1963) classic impossibility theorem in social choice theory.[12] The result is more general than Arrow's original theorem, since it is applicable not just to sets of propositions of the form $x \succcurlyeq y$, $y \succcurlyeq z$, and so on, but to sets of propositions with logical connections expressible in any well-behaved formal logic. We return to this point in Section 7.

6. Escape Routes

The following discussion of escape routes from the present impossibility builds on the analysis in Dietrich and List (2018), but focuses on a number of proposals that are specific to the context of ordinal moral uncertainty resolution.

6.1. Relaxing Universality

If our moral uncertainty resolution rule is not required to accept all possible credence functions on X as admissible inputs but is required to accept only credence functions that meet some further constraints, then the impossibility can be avoided.

A constraint that would trivially have this effect is the one that deems a credence function Cr on X admissible if and only if $f(Cr)$ is consistent and complete, where f is some antecedently fixed threshold rule. For instance, let f be the more-likely-than-not rule, under which, for each Cr, $f(Cr) = \{\, p \in X : Cr(p) > 0.5\}$ .If the domain of f is defined in terms of the given admissibility criterion (consisting of those credence functions Cr for which $f(Cr)$ is consistent and complete), then f clearly satisfies ordering, and, being a threshold rule, it satisfies certainty preservation and propositionwise independence too.

[12] Without the anonymity constraint on aggregation that is built into the present framework (by admitting as input only anonymous profiles of judgments, formally representable by credence functions), the four conditions *can* be simultaneously satisfied, but only by a dictatorial aggregation rule, which always produces the judgments of some antecedently fixed individual as output (Dietrich and List 2007, Dokow and Holzman 2010).

A slightly less trivial constraint that would suffice for consistent (albeit not generally complete) appropriateness judgments under a suitable threshold rule is the one that requires that, for each minimal inconsistent subset $Y \subseteq X$, there is at least one proposition $p \in Y$ for which $Cr(p) \leq t$. (A set of propositions Y is minimal inconsistent if it is inconsistent but every proper subset of Y is consistent.) A threshold rule with a strict threshold of t would never generate an inconsistent judgment set J for any credence function Cr satisfying that constraint, because no minimal inconsistent set of propositions Y would ever simultaneously get accepted under this threshold rule.

Incompleteness of J could be ruled out by further excluding credence functions that assign credences in the interval between $1 - t$ and t to any propositions in X. Putting this together would yield a moral uncertainty resolution rule on a restricted domain of admissible inputs, where that rule satisfies ordering, certainty preservation, and propositionwise independence, but of course not universality. Note, however, that this domain restriction would require some degree of decisiveness from the agent, as it requires that, for every proposition p, either $Cr(p)$ or $Cr(\neg p)$ exceed t. So, the agent would have to have a sufficiently strong opinion regarding p. Less confident agents will not satify this condition. Therefore, the present escape route from the impossibility offers a decision rule only for sufficiently decisive agents, where the required degree of decisiveness is given by t.

6.2. Relaxing Ordering

One natural way of relaxing ordering would be to require only the consistency and deductive closedness of the appropriateness judgment set J, not its completeness. This would correspond to generating a partial order $\succcurlyeq$ of the alternatives in A (a reflexive and transitive but possibly incomplete binary relation). So, we would replace the ordering condition with the following weaker condition:

Partial ordering. The output of f is always a consistent and deductively closed judgment set $J \subseteq X$, or equivalently a partial order of the alternatives in A.

This route may be initially appealing to those who take appropriateness in the presence of moral uncertainty to be at least sometimes indeterminate.[13] On such a view, in some cases, there is just no fact of the matter regarding what one ought to do given one's moral uncertainty, and therefore an adequate moral uncertainty resolution rule should be expected to leave some appropriateness judgments unsettled, and to yield an incomplete appropriateness relation. Unfortunately, however, this relaxation would not get one very far, as implied by the second theorem in Dietrich and List (2018).

Theorem 1.2: If $|A| \geq 3$ (and regardless of whether X is $\mathcal{P}(\Omega)$ or $\mathcal{P}_{binary}$), the only moral uncertainty resolution rule satisfying universality, partial ordering, certainty preservation, and propositionwise independence is a threshold rule with a weak threshold of 1.

Such a moral uncertainty resolution rule would not really help "resolve" any moral uncertainty at all. Only those moral propositions $p \in X$ to which the agent assigns a credence of 1 would be deemed approriate. In other words, if consistency, deductive closedness and the other conditions are preserved, there appears to be no room to accommodate limited indeterminacy of appropriateness. Once one permits some indeterminacy (by allowing for some incompleteness in the output of the resolution rule),

[13] Hedden (2016b) argues for general scepticism regarding the "ought" of normative uncertainty, but some of his arguments may convince one to adopt more limited scepticism. A proponent of such limited scepticism might find the relaxation of completeness appealing. See Riedener (2021: Ch. 6) for a theory that accommodates incompleteness of appropriateness in the case of cardinal moral uncertainty.

one gets sweeping indeterminacy in all propositions about which the agent is less than certain.

To obtain more substantial possibilities, we would have to relax the ordering condition further, namely to:

> **(Suzumura) consistency**. The output of f is always a consistent judgment set $J \subseteq X$, or equivalently, a Suzumura consistent binary relation over the alternatives in A.

One can then show that, for a sufficiently high threshold, a threshold rule will always generate a consistent judgment set, while also satisfying universality, certainty preservation, and propositionwise independence. Specifically, let k be the size of the largest minimal inconsistent subset of X. For instance, if $X = \mathcal{P}_{binary}$, then $k = |A| + 1$, insofar as the largest minimal inconsistent subset of $\mathcal{P}_{binary}$ is of the form:

$$\{x_1 \succcurlyeq x_2, x_2 \succcurlyeq x_3, \ldots, x_{k-1} \succcurlyeq x_1, \neg x_1 \succcurlyeq x_{k-1}\},$$

where $x_1, x_2, \ldots, x_{k-1}$ are the alternatives in A. Then a threshold rule with a strict threshold of $\frac{k-1}{k}$ always generates a consistent judgment set J, while also satisfying universality, certainty preservation, and propositionwise independence. The reason is that if a credence greater than $\frac{k-1}{k}$ is required for the acceptance of any proposition, it is never possible for the agent to deem k distinct propositions acceptable when the agent is rationally committed to assigning a joint credence of 0 to them. That, in turn, means that, since the agent's credence function is probabilistically coherent, the agent will never deem an inconsistent set of k or fewer propositions acceptable. And so, if k is the size of the largest minimal inconsistent subset of X, the agent will never end up with an inconsistent judgment set J.

In sum, any moral uncertainty resolution rule that allows for some incompleteness, but also settles some moral uncertainty while satisfying consistency and the other conditions, will sometimes produce appropriateness relations that rank x weakly above y and y weakly above z but are silent about the relative appropriateness of x and z.[14] This may be disappointing if one was hoping to accommodate indeterminacy of appropriateness by relaxing completeness, since in cases like the above, one would arguably expect appropriateness among x and z to be determinate. Apparently, given universality, certainty preservation, propositionwise independence and consistency, one may only accommodate indeterminacy by either accepting sweeping indeterminacy (with respect to any proposition one is even slightly uncertain about) or accepting that appropriateness among x and z may be indeterminate even when $x \succcurlyeq y$ and $y \succcurlyeq z$ are both determinate. Arguably, both options involve more indeterminacy than one would hope for. So, given the other conditions, it is impossible to be a moderate regarding indeterminacy of appropriateness: it is impossible to introduce *some* indeterminacy without ending up with (arguably) too much of it.

6.3. Relaxing Certainty Preservation

Certainty preservation is a very undemanding and compelling condition, so that it is hard to make a case for relaxing it if one wishes to take the agent's uncertainty into account. However, not all views wish to do so. Prominently, a normative externalist like Weatherson (2014, 2019) would argue that what an agent ought to do is independent of their moral uncertainty. So, the agent ought to choose x over y if and only if x is *in fact* at least as choice-worthy as y. Therefore, even if the agent's credence function is uncertainty-free, and they are maximally confident that a certain moral theory is true, they need not, according to the externalist, choose according to that theory. Rather, they ought to choose

[14] That is, such a rule will sometimes produce intransitive albeit Suzumura consistent appropriateness relations. Recall that a relation is Suzumura consistent if it is free from any pairwise comparison cycles in which at least one comparison is strict.

according to the true moral theory, i.e., the true choice-worthiness ranking. Interestingly, if $X = \mathcal{P}(\Omega)$, the only way to relax certainty preservation while leaving the other conditions intact is to adopt some version of normative externalism, i.e., some view that holds that what one ought to do is independent of one's moral uncertainty. To see this, consider the following result (from Theorem 5 in Dietrich and List 2018):

> **Theorem 1.3**: If $|A| \geq 3$ and $X = \mathcal{P}(\Omega)$, the only moral uncertainty resolution rules satisfying universality, ordering, and propositionwise independence are the constant rules, where there exists an antecedently fixed appropriateness judgment set J (consistent and complete) such that $f(Cr) = J$ for all credence functions Cr.

So, when $X = \mathcal{P}(\Omega)$, the class of normative externalist decision rules is fully characterized by the conjunction of universality, ordering, and propositionwise independence.[15]

6.4. Relaxing Propositionwise Independence

The limitations of the other escape routes from our impossibility suggest that, as in the context of belief binarization more generally, relaxing propositionwise independence offers the most plausible (or least implausible) escape route. When we allow the agent's all-things-considered judgment on a moral proposition to depend on their credences not just in that proposition itself, but also in other propositions, then we can avoid the present impossibility. Indeed, most moral uncertainty resolution rules proposed in the literature require us to give up propositionwise independence.

[15] We use the term "decision rule" instead of "moral uncertainty resolution rule" because the normative externalist is not really in the business of resolving moral uncertainty, despite making claims about how one ought to decide.

A prominent example of such a rule is the *Borda rule*, proposed by MacAskill (2016) in analogy to the equally named voting rule in social choice theory. To introduce it, we must first define the notion of a *Borda score* for an alternative at a given moral world. Recall that each moral world in Ω is a complete order $\succcurlyeq$ of the alternatives in A. At each such world $\succcurlyeq$, let the *Borda score* of alternative $x \in A$ be the number of alternatives $z \in A$ that are ranked at least weakly below x by the given order $\succcurlyeq$, formally $B_{\succcurlyeq}(x) = |\{ z \in A: x \succcurlyeq z\} |$.[16] If the agent's credence function Cr is defined on $X = \mathcal{P}(\Omega)$, then we can define the total Borda score for any alternative $x \in A$ as the credence-weighted average of its Borda scores at different worlds (orders) in Ω, formally,

$$B_{Cr}(x) = \sum_{\succcurlyeq \in \Omega} [Cr(\succcurlyeq) B_{\succcurlyeq}(x)] = \sum_{\succcurlyeq \in \Omega} [Cr(\succcurlyeq) |\{z \in A: x \succcurlyeq z\}|].$$

Here, $Cr(\succcurlyeq)$ is a shorthand for $Cr(\{\succcurlyeq\})$. The all-things-considered appropriateness judgment set $J = f(Cr)$ is then defined as the unique consistent and complete subset of X such that, for any alternatives $x, y \in A$,

$$(x \succcurlyeq y) \in J \text{ if and only if } B_{Cr}(x) \geq B_{Cr}(y).$$

So far, this definition only works when Cr is defined on $X = \mathcal{P}(\Omega)$, or specifically when credences are assigned to all of the worlds (orders) in Ω. However, we can also restate the definition for the case in which Cr is defined only on $X = \mathcal{P}_{binary}$, i.e., for propositions of the form $x \succcurlyeq y$ and their negations. To do so, note that

$$\sum_{\succcurlyeq \in \Omega} [Cr(\succcurlyeq) |\{z \in A: x \succcurlyeq z\}|] = \sum_{z \in \mathrm{A}} Cr(x \succcurlyeq z),$$

and so, the total Borda score for any alternative $x \in A$ can be equivalently defined as:

[16] An alternative variant of this definition would define the Borda score of alternative $x \in A$ as the number of alternatives $z \in A$ that are ranked strictly (rather than weakly) below x by $\succcurlyeq$, formally $|\{ z \in A: x \succ z\}|$. Both variants of the definition, which differ subtly in how they treat ties between alternatives, can be traced back to a corresponding Borda voting rule in social choice theory.

$$B_{Cr}(x) = \sum_{z \in \mathrm{A}} Cr(x \succcurlyeq z).$$

This is well-defined irrespective of whether $X = \mathcal{P}(\Omega)$ or $X = \mathcal{P}_{binary}$. We can then define the appropriateness judgment set under the Borda rule, $J = f(Cr)$, as the unique consistent and complete subset of X such that, for any alternatives $x, y \in A$,

$$(x \succcurlyeq y) \in J \text{ if and only if } \sum_{z \in \mathrm{A}} Cr(x \succcurlyeq z) \geq \sum_{z \in \mathrm{A}} Cr(y \succcurlyeq z).$$

It is evident that the Borda rule violates propositionwise independence. Whether x is deemed at least as appropriate as y depends not merely on the agent's credence in the proposition $x \succcurlyeq y$, but also on their credences in the propositions $x \succcurlyeq z$ and $y \succcurlyeq z$ for all other alternatives $z \in A$. On the other hand, the Borda rule clearly satisfies universality, ordering, and certainty preservation.

Another recent proposal that is inspired by ideas from social choice theory is Tarsney’s (2019) *McKelvey rule*. It is based on a binary relation on the set A of alternatives called the *McKelvey covering relation*. Given a credence function Cr, we say that alternative $x \in A$ *McKelvey-covers* alternative $y \in A$ if three conditions hold:

(1) $Cr(x \succcurlyeq y) > Cr(y \succcurlyeq x)$;

(2) For all $z \in A$, if $Cr(z \succcurlyeq x) > Cr(x \succcurlyeq z)$, then $Cr(z \succcurlyeq y) > Cr(y \succcurlyeq z)$; and

(3) For all $z \in A$, if $Cr(z \succcurlyeq x) \geq Cr(x \succcurlyeq z)$, then $Cr(z \succcurlyeq y) \geq Cr(y \succcurlyeq z)$.

The first condition says that x is “strictly credally preferred” to y, in the sense that the agent assigns a higher credence to x being at least as choice-worthy as y than to y being at least as choice-worthy as x. The second condition says that any alternative z that is strictly credally preferred to x is also strictly credally preferred to y. And the third condition says that any alternative z that is weakly credally preferred to x is also weakly credally preferred to y. The social-choice-theoretic precursor of this definition replaces "credally preferred" with "majority preferred". It is easy to see that the McKelvey covering relation on the set A is transitive and asymmetric. Importantly, however, it is incomplete: for some pairs of alternatives x and y (including the special case $x = y$), neither of the two alternatives McKelvey-covers the other. Tarsney (2019) argues that, in the face of moral uncertainty, an agent should choose a McKelvey-uncovered alternative, i.e., an alternative that is not McKelvey-covered by any other alternative. Thus defined, the McKelvey rule is not yet a moral uncertainty resolution rule in our sense, as it provides only a choice recommendation rather than a full appropriateness relation over the set of alternatives. But the rule may be extended to a full moral uncertainty resolution rule in various ways. One natural way to do so is as follows. For any credence function Cr, let the all-things-considered appropriateness judgment set $J = f(Cr)$ be defined as the unique, consistent, and complete subset of X such that, for any alternatives $x, y \in A$,

$(x \succcurlyeq y) \in J$ if and only if x is McKelvey-uncovered or y is McKelvey-covered.

The appropriateness relation yielded by this rule deems all uncovered options to be equally appropriate, all covered options to be equally appropriate, and all uncovered options to be strictly more appropriate than all covered options. While the rule satisfies universality and ordering, it violates propositionwise independence, since the criterion for whether a proposition such as $x \succcurlyeq y$ is included in J, namely whether x is uncovered or y is covered, depends not only on the agent's credence in the pairwise comparison between them, but also on their credences in pairwise comparisons involving other alternatives.[17] In its present form, the rule further violates certainty preservation, because whenever the agent is certain of a particular moral ranking (i.e., they have an uncertainty-

[17] The rule's satisfaction of universality is obvious. To see that the rule also satisfies ordering, it suffices to verify that the appropriateness relation encoded by $J = f(Cr)$ is always complete and transitive. To verify completeness, note that every option is either McKelvey-covered or McKelvey-uncovered. To verify transitivity, suppose that $x \succcurlyeq y \in J$ and $y \succcurlyeq z \in J$. So, x is uncovered, or y is covered and therefore z is covered. In consequence, x is uncovered or z is covered (or both), and in either case $x \succcurlyeq z \in J$.

free credence function), all non-maximal elements of that ranking are McKelvey-covered and will therefore be deemed equally appropriate by the rule even if they are not equally choice-worthy according to the ranking of which the agent is certain. However, the rule may be amended to satisfy certainty preservation. If Cr is uncertainty-free, then we may simply define $J = f(Cr)$ as the set of all propositions $p \in X$ such that $Cr(p) = 1$, while if Cr is not uncertainty-free, then we continue to use the definition above, i.e., $J = f(Cr)$ is the unique consistent and complete subset of X such that, for any alternatives $x, y \in A$,

$(x \succcurlyeq y) \in J$ if and only if x is McKelvey-uncovered or y is McKelvey-covered.

This amended rule satisfies universality, ordering, and certainty preservation, and only violates propositionwise independence.

One might also employ the McKelvey covering relation in a different way to obtain a moral uncertainty resolution rule, namely by defining $J = f(Cr)$ for a given credence function Cr as the smallest deductively closed (rather than complete) subset of X such that, for any alternatives $x, y \in A$,

$(x \succcurlyeq y) \in J$ if and only if x McKelvey-covers y.

This alternative McKelvey rule, however, would satisfy only partial ordering and potentially leave many pairwise comparisons between alternatives indeterminate.[18]

A third proposal from the literature that violates propositionwise independence is "My Favourite Theory", henceforth called the MFT rule, which requires a morally uncertain agent to choose in accordance with the theory that they have most credence

[18] Indeed, Brandt et al. (2016) show that the McKelvey covering relation can be any asymmetric and transitive relation (that is, for any asymmetric and transitive relation R over A, there is a credence function such that R is the McKelvey covering relation for that credence function). Therefore, the appropriateness relations yielded by this rule can be indeterminate to various degrees: from a copmlete strict order of the alternatives to the empty relation which is silent about appropriateness altogether.

in.[19] Specifically, given a credence function Cr, which is here assumed to be defined on $X = \mathcal{P}(\Omega)$, this proposal mandates that the agent should identify a world (order) $\succcurlyeq \in \Omega$ in which they have maximal credence (i.e., such that $Cr(\succcurlyeq) \geq Cr(\succcurlyeq')$ for all $\succcurlyeq' \in \Omega$) and then accept all and only those moral propositions $p \in X$ that are true at that world (i.e., such that $\succcurlyeq \in p$). Formally, $f(Cr)$ is a consistent and complete subset J of X such that, for any $\succcurlyeq \in \Omega$,

$$\{\succcurlyeq\} \in J \text{ only if } Cr(\succcurlyeq) \geq Cr(\succcurlyeq') \text{ for all } \succcurlyeq' \in \Omega.$$

A credence-maximizing world (order) $\succcurlyeq \in \Omega$ is interpreted as corresponding to an agent's “favourite theory”. If that world $\succcurlyeq \in \Omega$ is non-unique, then the definition requires a tie-breaking rule for choosing a “favourite” among them.[20]

The MFT rule satisfies universality, ordering, and certainty preservation, but violates propositionwise independence. To see this, note that two different credence functions may assign the same credence to a moral proposition p but render different

[19] See Gracely (1996b), Gustafsson and Torpman (2014), and Gustafsson (2022b).

[20] As discussed in Gustafsson and Torpman (2014), one could also adopt a different approach to breaking ties between credence-maximizing theories. Specifically, they argue, when more than one theory is maximally likely for the agent, this agent is permitted to choose in accordance with any of their maximally likely theories. So, one might stipulate that, for any credence function Cr and any two alternatives $x, y \in A$, $x \succcurlyeq y \in f(Cr)$ if and only if $x \succcurlyeq y$ holds at some world (order) $\succcurlyeq \in \Omega$ such that $Cr(\succcurlyeq) \geq Cr(\succcurlyeq')$ for all $\succcurlyeq' \in \Omega$. Unfortunately, however, this definition can easily yield an inconsistent appropriateness judgment set. For example, suppose the agent has positive credence in only two worlds (orders): $\succcurlyeq_1$, which places y above z above x, and $\succcurlyeq_2$, which places z above x above y, and $Cr(\succcurlyeq_1) = Cr(\succcurlyeq_2) = 0.5$, so that $\succcurlyeq_1$ and $\succcurlyeq_2$ are equally likely for her. According to the amended version of the MFT rule, $x \succcurlyeq y \in J$ (as this proposition is true at $\succcurlyeq_2$), $y \succcurlyeq z \in J$ (as this proposition is true at $\succcurlyeq_1$), but $x \succcurlyeq z \notin J$ (as this proposition is false at both $\succcurlyeq_1$ and $\succcurlyeq_2$), rendering the resulting appropriateness relation intransitive. Moreover, $z \succ x \in J$ (as $z \succcurlyeq x$ is true at both worlds while $x \succcurlyeq z$ isn't), and thus the resulting appropriateness relation is Suzumura inconsistent. Gustafsson and Torpman (2014) are aware of this problem and suggest that transitivity may be preserved by prohibiting the agent from choosing according to a theory that they violated recently, even if such a choice is prescribed by one of the theories they consider maximally likely (p. 168). So, in our example, if the agent chose x over y in accordance with $\succcurlyeq_2$ and then the agent is presented with a choice between y and z, they should not choose y as prescribed by $\succcurlyeq_1$ because they recently violated that theory in the choice of x over y. While this proposal might circumvent the identified problem in certain concrete situations of choice, it is arguably *ad hoc* and renders appropriateness judgments sensitive to the order in which alternatives are presented to the agent.

worlds (orders) maximally likely in a way that makes a difference for the acceptance of that proposition. For example, suppose the agent distributes their credence among three worlds: $\succcurlyeq_1$, which places x above y above z; $\succcurlyeq_2$, which places y above x above z; and $\succcurlyeq_3$, which places z above y above x. Let Cr_1 and Cr_2 be the following credence functions: $Cr_1(\succcurlyeq_1) = 0.4$, $Cr_1(\succcurlyeq_2) = Cr_1(\succcurlyeq_3) = 0.3$, and $Cr_2(\succcurlyeq_1) = 0.4$, $Cr_2(\succcurlyeq_2) = 0.5$, $Cr_2(\succcurlyeq) = 0.1$. Then Cr_1 and Cr_2 assign the same credence of 0.4 to the proposition $x \succ y$, but Cr_1 renders $\succcurlyeq_1$ maximally likely while Cr_2 renders $\succcurlyeq_2$ maximally likely, and so, according to the MFT rule, $(x \succcurlyeq y) \in J_1$ while $(x \succcurlyeq y) \notin J_2$, in violation of propositionwise independence.

A related proposal is "My Favourite Option", henceforth called the *MFO rule*.[21] According to this proposal, the agent should choose an alternative $x \in A$ for which they have the highest credence in its being most choice-worthy. As in the case of the original McKelvey rule, the MFO rule yields only a choice prescription rather than a full appropriateness relation over the alternatives in A. However, the definition may be extended to a full moral uncertainty resolution rule. To offer a natural such extension, define the set F_{Cr} of an agent's favourite alternatives for credence function Cr as the set of all $x \in A$ for which the agent has maximal credence in the proposition "x is most choice-worthy". More formally,

$$F_{Cr} = \left\{ x \in A \colon Cr\left(\bigcap_{y \in A} x \succcurlyeq y \right) \geq Cr\left(\bigcap_{y \in A} z \succcurlyeq y \right) \text{ for all } z \in A \right\}.$$

The MFO rule would then yield an appropriateness relation according to which all favourite alternatives are equally appropriate, all non-favourite alternatives are equally appropriate, and all favourite alternatives are strictly more appropriate than all non-favourite ones. Technically, the all-things-considered appropriateness judgment set

[21] This proposal is discussed, but not endorsed, by Lockhart (2000b), Gustafsson & Torpman (2014b) and Gustafsson (2022b).

$J = f(Cr)$ is defined as the unique consistent and complete subset of X such that, for any alternatives $x, y \in A$,

$$(x \succcurlyeq y) \in J \text{ if and only if } x \in F_{Cr} \text{ or } y \notin F_{Cr}.$$

This rule violates propositionwise independence for reasons similar to those for which the MFT rule does so: two different credence functions may assign the same credence to a given moral proposition but render different alternatives favourite in a way that makes a difference for the acceptance of that proposition by the MFO rule.[22] The rule satisfies universality and ordering.[23] By contrast, the MFO rule violates certainty preservation. For example, if $Cr(x \succ y) = Cr(y \succ z) = Cr(x \succ z) = 1$, then the rule will yield the judgments $x \succ y$, $x \succ z$, and $y \approx z$, because y and z are non-favourite. However, the definition of the MFO rule may be trivially amended in the same way in which we amended the McKelvey rule earlier so as to avoid this violation.

For a final example of a proposal that becomes possible if we give up propositionwise independence, again inspired by ideas from social choice theory, we introduce the class of *distance-based rules*. (These are also briefly mentioned by MacAskill (2016), though not defined in full generality.) Let a *distance metric* be a function d that assigns to each pair of an appropriateness judgment set $J \subseteq X$ and a credence function Cr on X a non-negative number $d(J, Cr)$ interpreted as the "distance" between J and Cr,

[22] For example, assume the agent distributes her credence among three moral worlds (orders): $\succcurlyeq_1$, at which $x \succ y \succ z$; $\succcurlyeq_2$, at which $y \succ x \succ z$; and $\succcurlyeq_3$, at which $z \succ y \succ x$. And let Cr_1 and Cr_2 be the following credence functions: $Cr_1(\succcurlyeq_1) = 0.4$, $Cr_1(\succcurlyeq_2) = Cr_1(\succcurlyeq_3) = 0.3$, and $Cr_2(\succcurlyeq_1) = 0.4$, $Cr_2(\succcurlyeq_2) = 0.5$, $Cr_2(\succcurlyeq_3) = 0.1$. Then while Cr_1 and Cr_2 agree on the proposition $x \succ y$ ($Cr_1(x \succ y) = Cr_2(x \succ y) = 0.4$), Cr_1 renders x uniquely favourite and Cr_2 renders y uniquely favourite, and therefore according to the MFO rule, $x \succ y \in J_1$ and $x \succ y \notin J_2$, in violation of propositionwise independence.

[23] In particular, to see that the generated appropriateness relation is complete, note that every alternative is either favourite or non-favourite, and to see that it is transitive, note that if $x \succcurlyeq y \in J$ and $y \succcurlyeq z \in J$ then $x \in F$ or $z \notin F$, and in either case $x \succcurlyeq z \in J$. So, for every credence function Cr, $J = f(Cr)$ is a well-defined complete and consistent subset of X.

subject to one constraint: the distance between a judgment set J and a credence function Cr is zero if and only if Cr is uncertainty-free and matches precisely the judgments in J, formally

> $d(J, Cr) = 0$ if and only if Cr coincides with J's *membership function*, defined as follows: for all $p \in X$,
>
> $$J(p) = \begin{cases} 1, & \text{if } p \in J; \\ 0, & \text{if } p \notin J. \end{cases}$$

A *distance-based moral uncertainty resolution rule* now assigns to each credence function Cr a consistent and complete judgment set $J \subseteq X$ for which $d(J, Cr)$ is minimal. (If there is no unique such J with minimal distance from Cr, this definition requires some tie-breaking criterion.) One simple distance metric is the "Hamming distance". Here, for any judgment set J and any credence function Cr, we have:

$$d(J, Cr) = \sum_{p \in X} |J(p) - Cr(p)|.$$

The judgment set J assigned to each credence function Cr will then be a consistent and complete subset of X that minimizes the distance from Cr thus defined.

A distance-based moral uncertainty resolution rule clearly satisfies universality, ordering, and certainty preservation (in light of the proviso giving necessary and sufficient conditions for $d(J, Cr) = 0$), but it obviously violates propositionwise independence. It displays a certain form of "holism", making the all-things-considered judgment on each moral proposition dependent on the credences in an entire "web" of other moral propositions.

The social-choice-theoretic precursor of the present definition, which requires reinterpreting Cr as encoding an anonymous profile of individual judgments or opinions across a group, as explained earlier, is called the *Hamming rule* (e.g., Pigozzi 2006) or, in the special case $X = \mathcal{P}_{Binary}$, the *Kemeny rule*. The Kemeny rule has the further property that it generates as its output the majority preference relation on the set of alternatives whenever that relation is transitive (so that we have an ordering), and it deviates from that output only in case the majority preferences need to be revised to restore ordering. (The majority preference relation is the binary relation $\succcurlyeq$ which, for any pair of alternatives $x, y \in A$, ranks x above y if and only if a majority of the voters do so.)

7. Beyond Ordinal Moral Uncertainty

Although we have focused on the resolution of ordinal moral uncertainty, we would like to conclude by indicating how our analysis can be extended to other forms of moral uncertainty. In general, the moral propositions about which an agent may be uncertain need not be restricted to propositions of the form $x \succcurlyeq y$ and their logical combinations. They could also include:

- propositions that are expressed by sentences with evaluative predicates, such as "x is good" or "x is virtuous", or
- propositions that are expressed by sentences with deontic operators, such as "it is permissible that p" or "it is obligatory that p".

There is no conceptual barrier to defining an algebra of propositions corresponding to such sentences, taken from a suitable predicate or deontic logic. The agent may assign credences to such propositions, for instance a credence of $2/3$ in the proposition "it is permissible that p", and may then have to decide whether or not to go ahead with some actions on that basis, for instance an action that would bring about p. Here the agent faces the question of what "accept"/"reject" verdicts ("appropriateness judgments”) to make on the given propositions, based on their credences in them. As before, the problem is an instance of the “belief-binarization problem”.

Formally, there is a non-empty set X of moral propositions of interest, called the *agenda*, and the agent has a credence function Cr on X, a real-valued function from X into [0,1] that is probabilistically coherent. The aim is to come up with a judgment set on X, i.e., a subset of X consisting of all those propositions in X that the agent “accepts” for the purposes of action guidance. A moral uncertainty resolution rule is defined as a function f that assigns to each credence function Cr on X (in some domain of admissible credence functions) a resulting appropriateness judgment set $J \subseteq X$.

Crucially, the framework on which our analysis is based imposes very few restrictions on the nature of the propositions that can be included in X (aside from the constraint that X must be closed under negation). We could, for instance, use any logic to express those propositions, provided the logic satisfies some minimal conditions. In particular, the logic must have a notion of negation and a notion of consistency, such that proposition-negation pairs are inconsistent, subsets of consistent sets of propositions remain consistent, the empty set of propositions is consistent, and any consistent set of propositions can be extended to a consistent superset containing a member of each proposition-negation pair within the logic. On these conditions, see Dietrich (2007). For instance, the propositions could be expressed in any standard propositional, predicate, modal, conditional, or deontic logic. We could then define the set Ω of all possible “moral worlds” as the set of all maximal consistent sets of propositions within that logic, and we would thereby be in a position to define an algebra on Ω and identify the agenda X, in the simplest case, with that algebra itself. The above-stated impossibility theorem (from Dietrich and List 2018) would immediately apply.[24] Of course, the condition we have

[24] This is under the non-triviality assumption that the algebra contains at least two distinct proposition-negation pairs beyond the pair Ω, $\emptyset$.

called "ordering" should now better be called "consistency and completeness of appropriateness judgments". Once again, we would be able to taxonomize possible moral uncertainty resolution rules in terms of which of the theorem's conditions they require giving up.

Even the case of cardinal moral uncertainty can in principle be accommodated within our framework. Here, we could introduce a set Ω of possible "moral worlds" that are each formally represented by a cardinal value function v, which assigns to each alternative x in A a real number $v(x)$ that numerically expresses the value of alternative x at that moral world. So, Ω becomes the set of all admissible cardinal value functions. An agent could then distribute their credence among such value functions. Furthermore, if we define an algebra on Ω – for instance, a set of measurable subsets of Ω – then we can express propositions about the alternatives' cardinal value or value comparisons. For example, the proposition "the value of x is greater than the value of y" would correspond to the subset of Ω consisting of those value functions for which $v(x) > v(y)$. Similarly, the proposition "a switch from x_1 to y_1 is twice as valuable as a switch from x_2 to y_2" would correspond to the subset of Ω consisting of those value functions for which $v(y_1) - v(x_1) = 2(v(y_2) - v(x_2))$. And if values were thought to be sufficiently precisely measurable, a proposition such as "the value of x exceeds a particular threshold" would correspond to the subset of Ω consisting of those value functions for which $v(x)$ is above the specified threshold. A credence function on the present algebra would capture an assignment of credences to such propositions.

A complication is that value functions need not be fully unique but are usually assumed to be unique only up to a certain class of transformations. For instance, if value is measurable on an interval scale, then value functions are unique only up to *positive affine transformations*, which means that any value function v' which can be expressed as

$v' = av + b$, for some positive real number a and real number b, would be deemed equivalent to v. That is, any such v and v' would be deemed to have the same informational content. Similarly, if value is merely ordinally measurable, then value functions are unique only up to positive monotonic transformations, which means that any two value functions v and v' which induce the same ordering over the alternatives in A (i.e., for which, for any x and y in A, $v(x) \geq v(y)$ if and only if $v'(x) \geq v'(y)$) would be deemed to have the same informational content.

To capture this lack of full uniqueness of value functions, one would need to say that the moral worlds aren't individual value functions, but only equivalence classes of value functions that are deemed to be informationally equivalent. In the case of interval-scale measurability, for instance, each moral world would consist of value functions that can be derived from one another via the formula $v' = av + b$ for some real numbers a and b with $a > 0$. The equivalence classes would be singletons only in the extreme special case in which value is perfectly measurable on some absolute scale. Still, the present way of defining Ω – as a set of equivalence classes of value functions, such that all the value functions within each equivalence class are assumed to represent the same "moral world" – would allow us to induce an algebra of moral propositions, to which our formal framework could once more be applied.

Interestingly, our impossibility theorem continues to hold even when the agenda of moral propositions under consideration is not a full algebra on Ω but a much smaller set of propositions. All that is needed for the theorem to apply is that the agenda X satisfies a richness constraint called "blockedness" (originally introduced in the context of judgment aggregation by Nehring and Puppe 2002). Roughly speaking, an agenda X is *blocked* if there exists at least one proposition in X for which there exists a "path of conditional entailments" from that proposition to its negation and another path in the opposite direction; for details, we refer readers to Dietrich and List (2018).[25] For example, a set consisting of two distinct "atomic" propositions p and q, their conjunction $p \cap q$, and their disjunction $p \cup q$, as well as the negations of all these propositions is blocked, and so the impossibility theorem applies. Another example is the set consisting of two distinct such propositions p and q, the material biconditional $p \leftrightarrow q$, and their negations.

These considerations suffice to illustrate that the impossibility of moral uncertainty resolution in accordance with universality, consistency and completeness of appropriateness judgments, certainty preservation, and propositionwise independence is quite pervasive. And even if consistency and completeness of appropriateness judgments were weakened to deductive closedness alone (the counterpart of partial ordering above), only a degenerate uncertainty resolution rule would become possible which deems a proposition acceptable if and only if the agent's credence in it is 1 (Dietrich and List 2018). In consequence, the most promising escape routes from the impossibility are the ones that abandon propositionwise independence and thereby take a more holistic approach to moral uncertainty resolution: one's appropriateness judgment on a proposition p will not generally depend only on one's credence in p, but may also depend on one's credences in related propositions.

The only other way in which we could avoid the impossibility without restricting the domain of admissible credence functions or giving up certainty preservation or consistency and completeness of appropriateness judgments would be to shrink the agenda of moral propositions so as to remove enough logical connections from it to render it no longer "blocked". For instance, if, for the purposes of action guidance, we

[25] Technically, a path of conditional entailments from p to q is a sequence of propositions $p_1, \dots, p_k \in X$ where $p_1 = p$ and $p_k = q$ such that, for each $i = 1, \dots, k-1$, p_i conditionally entails p_{i+1}. Here, we say that proposition $p \in X$ conditionally entails proposition $q \in X$ if there is some subset $Y \subseteq X$, consistent with each of p and the negation of q, such that $\{p\} \cup Y$ entails q.

could get away with only adjudicating a set of logically independent propositions, then the identified impossibility would no longer apply. Since this route, however, is likely to be very limited in applicability, we must conclude that the problem of moral uncertainty resolution – whether ordinal or not – has no obvious or privileged solution. There are solutions, as reviewed in this chapter, but they all come with some theoretical costs.

Chapter 2

Weak Norms for Moral Uncertainty

1. Introduction

Uncertainty about moral matters is ubiquitous. People considering vegetarianism may be uncertain how to morally trade off human and animal welfare, policy makers considering alcohol taxation might wonder whether paternalistic interventions are morally justifiable by the benefits they generate, and when donating to charity, one could be unsure whether moral obligations are stronger to members of one's community than to other people. In all such cases uncertainty regarding some moral hypotheses or moral theories may give rise to uncertainty about the moral status of the alternatives under consideration. These types of cases pose a decision-theoretic challenge – how should morally uncertain agents choose? A growing body of philosophical work is concerned with answering this question and suggesting norms for the morally uncertain agent.

Philosophers grappling with this question have realized that its answer might depend on the structure of the moral properties alluded to by the hypotheses or theories the agent is uncertain about. For example, the moral hypotheses might only allude to permissibility, merely partitioning available options into permissible and impermissible subsets. Alternatively, the agent's uncertainty may involve ordinal statements ordering pairs of alternatives according to which is morally better but remaining silent about degrees of betterness. Finally, the agent may entertain hypotheses that involve a cardinal notion of moral value, mapping options onto an interval scale, and giving rise to judgements about some options being *much* better than others or merely *slightly* better, and more generally licensing comparisons of differences in moral value. Since the richness of the information included in the agent's epistemic state varies considerably between these cases, we should expect to have different decision rules aggregating this information for each case.

This chapter focuses on the latter case of *cardinal moral uncertainty*, in which the moral theories the agent is uncertain about all concern a cardinal notion of moral value. Each such theory yields judgments about the ratios of value differences among some set of alternatives – judgments of the form: "the value difference between alternatives x and y is n times the value difference between alternatives z and w" – enabling talk of degrees of betterness among alternatives. Cardinal moral theories are represented by interval-scale-measurable value functions (from alternatives to numbers) that are unique only up to positive affine transformation: if v represents a cardinal moral theory, then so does $v' = av + b$ for any positive number a and any number b. This non-uniqueness of the numbers a cardinal moral theory assigns to the alternatives reflects the fact that such theories do not specify a unit of value or a zero point – the only features of a particular value function v that represent genuine properties of the moral theory are those that are invariant to changes in the unit or zero point. So, the case of cardinal moral uncertainty consists in an agent who divides their credence between multiple interval-scale value functions (unique up to positive affine transformation), each representing a competing hypothesis about the identity of the actual moral value function.

A natural response to the decision theoretic challenge posed by cardinal moral uncertainty is to apply ordinary decision theory and require, for example, that agents choose an alternative with maximal expected moral value relative to their credences. However, any such application would require settling comparisons of value differences across theories – settling judgments of the form: "the value difference between alternatives x and y on one moral theory is n times the value difference between alternatives z and w on another moral theory" – and at least in some cases, such comparisons are indeterminate. In some cases, for some pairs of value differences, it is neither true nor false that one is n times as great as the other. This problem of indeterminate *Intertheoretic Value Comparisons* (IVCs) adds an additional layer of complexity to the original problem of cardinal moral uncertainty and prevents the straightforward application of ordinary decision theory to solve it. The first level of the problem requires finding a norm for choice under moral uncertainty with no indeterminacy – let us call this the *lower-order* norm. The second layer of the problem requires finding a second norm for cases in which IVCs are indeterminate and the lower-order norm, which depends on IVCs, is indecisive. This second norm is a norm for choice under indeterminacy. So, even given some lower-order norm for cardinal moral uncertainty with no indeterminacy – e.g., an application of ordinary decision theory requiring the maximization of expected moral value – there is still the further question of how the agent should choose when IVCs are indeterminate and such a norm cannot be straightforwardly applied.

The problem of IVC indeterminacy has elicited two types of responses in the literature. First, some philosophers have proposed what I term *Strong norms* – norms that dissolve the indeterminacy and prescribe a set of alternatives that would be prescribed by ordinary decision theory given a particular way of settling IVCs. Some strong norms dissolve indeterminacy explicitly by picking out a particular way of settling IVCs while others arrive at their prescriptions using other mechanisms. Second, some philosophers respond to IVC indeterminacy by adopting general scepticism towards the normative significance of moral uncertainty. Such philosophers take IVC indeterminacy to be indicative of the intractability and perhaps even the unintelligibility of the problem of moral uncertainty.

However, examining the philosophical discussion of choice under indeterminacy more generally, reveals a third type of response that the moral uncertainty debate has overlooked. Decision theories for choice under indeterminacy are dominated by two types of norms: *Strong norms* and *Weak norms*. Strong norms, like the ones proposed for IVC indeterminacy, prescribe alternatives that would be prescribed by the lower-order norm given a particular way of settling the indeterminate facts. In contrast, weak norms are much less decisive – they prescribe all alternatives that would be prescribed by the lower-order norm given *all* the possible ways of settling the indeterminate facts. While Strong norms dissolve the indeterminacy and privilege a certain way of settling the indeterminate facts, Weak norms retain the indeterminacy and do not make distinctions among the different possible ways of settling it.

I have two goals in this chapter. First, I wish to call attention to Weak norms for moral uncertainty with IVC indeterminacy. If weak norms are a central approach to choice under indeterminacy, then, at least *prima fascia*, they should figure in prominently as a response to IVC indeterminacy. Without an argument for a fundamental difference between IVC indeterminacy and other types of indeterminacy – and I am unaware of such an argument – it is hard to see what justifies the dissimilarity between the general debate of choice under indeterminacy and the particular case of IVC indeterminacy in the context of moral uncertainty. I do not intend to argue for weak norms for moral uncertainty, but to call attention to them, and to argue that they ought to be considered seriously. This chapter may thus be seen as an invitation to consider Weak norms for moral uncertainty.

My second goal is to take up my invitation and begin considering a particular weak norm, for a special case of moral uncertainty I term *the expectational case*. The expectational case is characterized by two assumptions. First, that the moral theories the agent considers possible are expectational – the value of an alternative on each theory is its expected value. And second, that the lower-order norm for moral uncertainty with no indeterminacy is Expected Value Maximization, where the expected value of an alternative is the credence-weighted average of its values on the different moral theories the agent considers possible. I show that a central Weak norm for the expectational case is equivalent to the norm requiring the agent to choose non-dominated alternatives. I then use this equivalence to explore and evaluate the landscape of strong norms for the expectational case.

The chapter proceeds as follows. In section 2, I present and discuss the problem of indeterminacy, and the other problems posed by Intertheoretic value Comparisons. In section 3, I present a general framework for choice under indeterminacy and partition the norms for such cases into Weak and Strong camps. In section 4, I apply this framework and partition to moral uncertainty with indeterminate IVCs. In section 5, I consider a Weak norm for the expectational case and show that it is equivalent to non-dominance. In section 6, in light of the equivalence, I consider the landscape of strong norms for the expectational case. In section 7, I consider the *Bargaining norm* – a strong norm recently proposed in the literature. Finally, in section 8, I conclude.

2. Intertheoretic Value Comparisons and Indeterminacy

The problem of cardinal moral uncertainty may initially seem quite similar to the bread-and-butter problem of choice under ordinary, non-moral, uncertainty. In both cases the agent is uncertain about some hypotheses and the values of the alternatives at hand depend on which of these hypotheses is correct. The only difference between the cases, it might initially seem, is the content of the hypotheses the agent is uncertain about – in the moral case they concern morality, and in the ordinary case they concern other matters. Indeed, if this picture is correct, there is reason for optimism about decision theory for moral uncertainty, as the problem might be solvable by simply applying ordinary decision theory to the moral case.[1]

However, decision theory – in virtually all of its forms – requires comparing value differences across the hypotheses the agent considers possible, and such comparisons are importantly different in the moral and non-moral cases. To demonstrate this difference, and how it impedes the application of decision theory to the case of moral uncertainty, let us consider two toy examples – an ordinary, non-moral case, and then a moral case:

> **Danny** is uncertain whether the Hare Krishna society is serving free lunches at the university tomorrow and divides his credence between the event that they are, and the event that they aren't. He is deliberating whether to cook and bring his own lunch to university tomorrow. If free lunches were provided, Danny would rather not cook and eat the free lunch, but if there are no free lunches tomorrow, cooking and eating his own lunch would be preferable to no lunch at all. Danny's decision problem may be modelled using the following decision table, where a-d represent the values of the alternative actions under the relevant (non-moral) events, and p and $1 - p$ represents Danny's credences in those events:

TABLE 2.1: *Danny's decision problem*

	Free lunch (p)	No free lunch ($1 - p$)
Bring lunch	a	b
Don't bring lunch	c	d

Whether Danny should bring his lunch – on virtually any decision theory – depends on two types of comparisons. First, on a comparison of likelihoods: how likely the event of a free lunch is relative to the likelihood of its complement. The less likely the

[1] See MacAskill & Ord (2020b) who argue that, when possible, moral, and non-moral uncertainty should be treated analogously. See also Weatherson's (2019b) discussion of *Symmetry*.

event of a free lunch (as p decreases), the stronger the normative push in favour of Danny bringing his own lunch. As free lunches grow likelier (p increases), the normative push in Favor of going in lunchless grows stronger. Second, what Danny ought to do depends on a comparison of value differences: the difference in value between bringing lunch and going in lunchless given that there are no free lunches tomorrow ($b - d$), and the inverse value difference on the event of free lunches ($c - a$). The greater the first difference is relative to the second difference, the stronger the push to bring lunch. Different versions of decision theory will aggregate these normative forces in different ways, but all will make use of them to yield prescriptions.[2]

Now, compare Danny's case to Sally's:

> **Sally**, who lives in a developed country, is uncertain whether her moral obligations to members of her community are stronger than her moral obligations to other people. She may be thought of as dividing her credence between two cardinal moral theories: a communitarian theory according to which the moral value of an action is sensitive to whether the parties involved are members of the agent's community, and a universalist theory that treats such facts as morally insignificant. Sally is deliberating whether to donate money to a local food bank or to a famine relief charity operating in a remote famine-struck territory. Let us assume that on the communitarian theory, donating to the local food bank is morally better than donating to famine relief and on the universalist theory the inverse is true. The decision problem may be modelled using the following decision table, where e-f represent the values of the alternative actions under the relevant moral hypotheses, and q and $1 - q$ represents Sally's credences in those hypotheses:

[2] Most prominently, expected utility theory aggregates these factors as follows: it is permissible for Danny to bring his own lunch if and only if $\frac{b-d}{c-a} \geq \frac{p}{1-p}$. Other decision theories (e.g., Buchak 2013) have different aggregation procedures and might consider other factors in addition to these ratios, but their verdicts will always be sensitive to these ratios in the ways described in the text.

TABLE 2.2: *Sally's decision problem*

	Communitarian (q)	Universalist ($1-q$)
Donate to food bank	e	f
Donate to famine relief	g	h

Applying ordinary decision theory to Sally's case would require likelihood and value-difference comparisons. While the comparison of likelihoods seems straightforward – the agent's credence in the two theories determines the matter – the value-difference comparison is less obvious. Is the value difference between donating locally and donating to famine relief according to the communitarian theory greater than, smaller than, or equal to the inverse difference according to the universalist moral theory? If it is greater, is it much greater, or merely slightly greater? These questions must be answered for ordinary decision theory to be applicable to cases of moral uncertainty, and at least initially, it is not obvious how to answer them.

There are three main worries regarding comparisons of value differences across theories, which I will also call *Intertheoretic Value Comparisons (IVCs)*, that do not arise for corresponding comparisons in the non-moral case.[3] First, there's a worry about the meaning of such comparisons (Nissan-Rozen 2015). If the meaning of claims about moral value is cashed out in terms of properties of the value functions, then some claims will be rendered meaningless by the interval scale structure of these functions. For example, claims involving a single moral theory like "the value of x on moral theory V is 4" or "the value of x is twice the value of y on moral theory V" have no meaning when evaluated relative to an interval scale, as they are not robust with respect to positive affine

[3] Riedener (2021a) raises the first two of these problems, as well as a further worry, which I will not address here, about the type of facts that ground IVCs.

transformation and will therefore always be true on some value functions representing a cardinal theory and false on others.[4] Claims of this form would only be meaningful when evaluated relative to a richer background structure that fixes a unit and a zero point. Similarly, nontrivial comparisons of value differences involving two moral theories will never be robust to positive affine transformation of both theories.[5] So, nontrivial IVCs will always be rendered true by some pairs of value functions representing the two cited theories and rendered false by other pairs.

So, the argument goes, saying that the value difference between donating locally and donating to famine relief according to the communitarian theory is twice the inverse difference according to the universalist moral theory, is meaningless in the same way that saying that the value of donating to famine relief according to the communitarian theory is 4. Notice that this problem does not arise in the non-moral case, where the agent's desires are represented by a *single* utility function mapping all alternative-hypothesis pairs onto an interval scale, thereby fixing the value differences across hypotheses.

The worry of meaninglessness may, however, be alleviated if one understands IVCs as referring to some other properties, distinct from the moral theories and their associated value functions. For example, Ross (2006) and Riedener (2021a) adopt an explication of IVCs on which they refer to metanormative facts – facts about what one ought to do given one's moral uncertainty. Tarsney (2018a) offers an understanding of

[4] Or consider for example the claim "Jerusalem is twice as hot as London today" evaluated on a day in which it is 10 °C in London and 20 °C in Jerusalem. The ratio between 20 and 10 is simply not a property of the temperatures of the cities, rather, it is an artifact of the particular numerical representation of the Celsius scale, which is not preserved under positive affine transformation, e.g., to the Fahrenheit scale. So, such a claim is not true, but neither is it false as it is not the case that some other ratio holds between the temperatures of the two cities, rather, it is meaningless – it refers to properties that temperature doesn't possess. The claims in the text are meaningless in precisely the same way.

[5] Trivial IVCs are ordinal comparisons among value differences with different signs. So, when one value-difference is positive and another is negative or zero, the claim that the former is greater than the latter is a trivial IVC. Trivial IVCs are robust across positive affine transformation (as they track ordinal properties of the theories) and are therefore meaningful.

IVCs as claims referring to relations between points of agreement and disagreement between two moral theories. MacAskill et al. (2020) understand IVCs to be claims about moral theories but understand these theories as mapping alternatives onto the same "universal scale" rather than merely disparate sets of value functions. While contested, these explications all provide potential solutions to the problem of meaninglessness.

The second worry about IVCs, and the one I will focus on here, is regarding their determinacy. Even if some explication of the meaning of IVCs is accepted, and we know what it means for an IVC to be true, one might still worry that at least in some cases, there is just no fact of the matter as to whether an IVC is true or false. Indeed, the proponents of all three explications of IVCs above take them to be partially indeterminate. The worry of indeterminacy partially stems from the fact that at least in some cases, there seems to be no nonarbitrary way of settling IVCs. For example, it is notoriously difficult to come up with a good way of comparing value differences between average and total utilitarianism (Gracely 1996, Broome 2012, Hedden 2016, Cotton-Barratt et al. 2020). This problem has led some philosophers to accept that at least in some cases, value differences are "incomparable", or that comparisons are "impossible".[6] The problem is not that we don't know the conditions under which a comparison between two value differences would be true or false, rather, that neither condition – for truth nor falsity – appears to be satisfied.

At least conceptually speaking, indeterminacy of IVCs need not be an all-or-nothing matter. It is conceivable for example, that some IVCs are determinate – because the facts that constitute their truth conditions are determinate – while others are

[6] Hudson (1989), Gracely (1996a), Gustafsson & Torpman (2014a) take IVCs to be always indeterminate, Broome (2012), Tarsney (2018a), MacAskill et al. (2020), and Riedener (2021a) take them to be sometimes indeterminate. Although these philosophers don't use the term "indeterminacy" explicitly, it appears to be the best way to understand the problem they are referring to. The said "impossibility" of IVCs, does not appear to mean that IVCs are meaningless, as some of these authors are committed to certain explications of IVCs, and thus see themselves to have solved the problem of meaninglessness.

indeterminate. More generally, the scope of IVC indeterminacy may vary in two different ways – within comparisons and across comparisons. First, the degree of indeterminacy within a single intertheoretic value comparison may vary. For example, given two value differences (on different theories) d_1 and d_2, it may be that it is merely indeterminate whether d_1 is twice, or three times as great as d_2, or it may be that for any positive real number r it is indeterminate whether d_1 is r times d_2. Second, the degree of indeterminacy across comparisons may vary. For example, it may be the case that there is only indeterminacy when it comes to comparisons involving the value differences d_1 (on theory 1) and d_2 (on theory 2), but comparisons involving other value differences are fully determined. Or, in contrast, it could be the case that all value-difference comparisons involve some degree of indeterminacy (within comparisons).

IVC indeterminacy – whether limited or pervasive – prevents the simple application of ordinary decision theory to our case. As mentioned above, ordinary decision theory relies on the formal analogue of IVCs – comparisons of value differences across (non-moral) states. Therefore, providing norms for choice under moral uncertainty with indeterminate IVCs would require going beyond the resources of orthodox decision theory.

The third worry concerning IVCs is one of epistemic access. Even if the meaning of IVCs is settled, and they have determinate truth values, it still may be the case that the agent is uncertain about, or wholly ignorant of, these truth values. Such an agent may perhaps be represented by a credence function on some space of possible ways of settling IVCs, or a set of such credence functions. It is unobvious how we are to aggregate the agent's precise or imprecise credences about IVCs, their credences about moral theories, and the content of those moral theories. As far as I know this problem has not received attention in the literature.

Here I focus on the second problem, that of indeterminacy. Before analysing the problem further, it would be useful to represent it more formally. The formal framework I employ to do so has three basic components: a set of alternatives for moral evaluation, a set of cardinal moral theories that evaluate the alternatives, and the agent's credence distribution over those theories. First, let $A = \{a_1, \dots, a_m\}$ be a nonempty set of m *alternatives* – available options or courses of action evaluated by the moral theories. Second, let $\mathbb{V} = \{V_1, \dots V_n\}$ be the set of the n *cardinal moral theories* the agent gives positive credence to, where each cardinal moral theory is represented by a set of value functions V whose elements all map the alternatives to real numbers. As mentioned above, since cardinal moral theories map alternatives onto an interval scale, the elements of V are all *positive affine transformations* of each other, and all positive affine transformations of the elements of V are also elements of V (i.e., V is *closed* under positive affine transformation). More formally V satisfies the following two conditions:

> TRANSFORMATION: for any $v, v' \in V$ there exists $a \in \mathbb{R}^+$ and $b \in \mathbb{R}$ such that $v'(\cdot) = av(\cdot) + b$.
>
> CLOSURE: if $v \in V$, $a \in \mathbb{R}^+$ and $b \in \mathbb{R}$ then $v' \in V$ where $v'(\cdot) = av(\cdot) + b$.

Finally, the agent's credences may be represented by an n-tuple of non-negative numbers that sum to one: $p = (p_1, \dots, p_n)$, where p_i is the agent's credence in theory V_i.[7]

The moral theories thus characterized do not determine non-trivial IVCs, as value-difference comparisons across theories vary depending on the individual value functions chosen to express them in.[8] The moral theories therefore leave open multiple ways of

[7] Strictly speaking, the agent's credence function must be defined over an algebra, which in our case could be the powerset of $\mathbb{V}$. I ignore the non-atomic elements of this powerset as they will not matter for our purposes here.

[8] For example, it might be that for the value functions $v_i \in V_i$ and $v_j \in V_j$ the following value difference holds: $v_i(a) - v_i(b) > v_j(c) - v_j(d) > 0$ for $a, b, c, d \in A$. But this inequality does not hold if we choose

settling IVCs. These different ways of settling IVCs left open by the theories, may be represented by sets of *value tuples*, where a value tuple $\boldsymbol{v}$ of the set of moral theories $\mathbb{V}$ is any n-tuple of value functions $\boldsymbol{v} = (v_1, \dots, v_n)$ such that $v_i \in V_i$ for $i = 1, \dots, n$. Let us denote the set of all such value tuples $T(\mathbb{V})$, and let $\sim$ be a binary relation among value tuples in $T(\mathbb{V})$ such that for any two value tuples $\boldsymbol{v}, \boldsymbol{v}' \in T(\mathbb{V})$, $\boldsymbol{v} \sim \boldsymbol{v}'$ iff $\boldsymbol{v}'$ is a *positive affine transformation* of $\boldsymbol{v}$. That is, if there exist $a \in \mathbb{R}^+$ and $b \in \mathbb{R}$ such that $\boldsymbol{v}' = (av_1(\cdot) + b, \dots, av_n(\cdot) + b)$, where $\boldsymbol{v} = (v_1, \dots, v_n)$.[9] Since $\sim$ is an equivalence relation, it partitions $T(\mathbb{V})$ into equivalence classes. Let this partition be denoted by $C(\mathbb{V}) = \{[\boldsymbol{v}] | \boldsymbol{v} \in T(\mathbb{V})\}$, where $[\boldsymbol{v}]$ is the equivalence class of $\boldsymbol{v}$ under $\sim$, and let each such equivalence class be termed a *calibration*. All value tuples in a given calibration settle IVCs in the same way – they all yield the same value-difference ratios – and any two tuples from different calibrations will disagree about at least some IVCs. Therefore, the set of calibrations represents the full range of ways, consistent with the moral theories, of settling IVCs.[10]

Each calibration represents a unique way of settling IVCs consistent with the moral theories the agent considers possible, and every such way is represented by some calibration. IVC indeterminacy may therefore be represented by a set of calibrations $C \subseteq C(\mathbb{V})$ where the elements of C represent the different ways of settling IVCs that are left open by the facts. Different degrees of IVC indeterminacy may therefore be expressed by

other value functions that represent the same theories. For example, let $v_j' = rv_j$ such that $r > \frac{v_i(a)-v_i(b)}{v_j(c)-v_j(d)}$. v_j' is a positive affine transformation of v_j and therefor by CLOSURE, $v_j' \in V_j$. But $v_j'(c) - v_j'(d) > v_i(a) - v_i(b) > 0$, reversing the comparison we started with. This procedure can be applied to any non-trivial IVC, and therefore, non-trivial IVCs are sensitive to the choice of value functions used to express them and are not determined by the moral theories.

[9] Notice that the positive affine transformation relation defined here is among value *tuples*, in contrast to the positive affine transformation relation among value *functions* discussed above.

[10] Notice that while each value tuple settles IVCs in a particular way (as it determines a precise ratio between every pair of value differences), some pairs of value tuples will agree about all IVCs (namely the pairs that are positive affine transformations of each other). Therefore, while $T(\mathbb{V})$ is too fine-grained to capture the notion of IVCs, $C(\mathbb{V})$ – which is essentially a coarsening of $T(\mathbb{V})$ – gives us the notion we are out to represent.

different sets of calibrations – different subsets of $C(\mathbb{V})$. When IVCs are fully determinate, the facts pick out a single calibration from $C(\mathbb{V})$. In contrast, the case of maximal IVC indeterminacy, where all ways of settling IVCs are left open by the facts, is representable by the full set of possible calibrations $C = C(\mathbb{V})$. In between these two extremes, different subsets of $C(\mathbb{V})$ may be used to represent cases in which IVCs are only partially indeterminate. While much of the literature assumes that when IVCs are indeterminate, they are maximally indeterminate, the framework I propose here does not presuppose this and can accommodate cases of partial indeterminacy.[11]

In sum, the problem of cardinal moral uncertainty with IVC indeterminacy may be represented by a tuple of the form $(A, \mathbb{V}, p, C)$ where A is a set of alternatives, $\mathbb{V}$ is a set of cardinal moral theories, p is the agent's credence tuple, and C is the set of calibrations left open by the facts, which includes more than one element as IVCs are at least partially indeterminate.

While, as mentioned above, this problem is unamenable to ordinary decision theory, other general decision-theoretic tools may still be applicable to it. In particular, if what one ought to do in cases of moral uncertainty depends on how IVCs are settled, and IVCs are indeterminate, then our problem is an instance of the more general category of choice under indeterminacy – cases in which what one ought to do is indeterminate. This problem of choice under indeterminacy is a general decision-theoretic challenge, which has received attention in the literature quite independently of moral uncertainty. Therefore, to address our problem of moral uncertainty with indeterminate IVCs, it would be best to first examine the general case of choice under indeterminacy, and subsequently try to apply lessons from the general case to ours. I turn to this in the next sections.

[11] An important exception is Riedener (2021a) who allows for a varying degree of IVC indeterminacy. Carr (2022) considers varying degrees of indeterminacy *within* moral theories.

3. Choosing Under Indeterminacy

If what the morally uncertain agent ought to do depends on how moral values compare across moral theories, and such comparisons are indeterminate, then our problem is an instance of choice under indeterminacy. Taking a closer look at this general category of choice under indeterminacy, which has received philosophical attention independent of moral uncertainty, should therefore shed some light on the latter.

A choice situation is said to be *under indeterminacy* if the normative status of at least some of the alternatives on offer is indeterminate. In such cases, the evaluation of alternatives by the norms at play – be those norms of rationality, morality, metanormativity or any other norms – depends on whether some indeterminate propositions are true or false.[12] Since indeterminate propositions are neither true nor false – their truth values are unsettled by the facts – the norms at play cannot arrive at a determinate evaluation of the alternatives in the way possible when the relevant propositions are true or false. A determinant evaluation is only possible given some *sharpening* of the indeterminant propositions – some consistent truth assignment settling their truth values. Choice under indeterminacy is thus characterized by some of the alternatives' normative status varying across different sharpenings, across different ways of settling the indeterminate propositions.[13]

To model this phenomenon, let A be some set of alternatives and let S be the set of sharpenings of some set of indeterminate propositions. Let us assume that the relevant norms tell us, when there is no indeterminacy, which elements of A are permissible. These norms may be represented by a function N from sharpenings to sets of alternatives

[12] Indeterminacy in the normative status of alternatives may also stem from indeterminacy in the norms themselves, rather than the facts that the norms take as inputs (see Schoenfield 2016). I set this case aside as it involves a different kind of indeterminacy than the one generated by the problem of IVCs.

[13] I adopt this characterization, and the term "sharpening" from Williams (2014).

(subsets of A). N assigns each sharpening with the set of alternatives that are permissible given that sharpening. Cases of choice under indeterminacy are precisely those in which the verdicts of N are not identical across all sharpenings, i.e., when there exist sharpenings $s, s' \in S$ such that $N(s) \neq N(s')$.

On this picture, alternatives may be partitioned into the following three categories with respect to their permissibility:

1. **Determinately permissible** alternatives are those that are permissible on all sharpenings: $a \in A$ is determinately permissible if for all $s \in S$, $a \in N(s)$.
2. **Determinately impermissible** alternatives are those that are impermissible on all sharpenings: $a \in A$ is determinately impermissible if for all $s \in S$, $a \notin N(s)$.
3. **Indeterminately permissible** alternatives are those that are permissible on some sharpenings and impermissible on others: $a \in A$ is indeterminately impermissible if there exist $s, s' \in S$, such that $a \in N(s)$ and $a \notin N(s')$.

The decision-theoretic challenge generated by indeterminacy stems from the unobvious relationship between indeterminate permissibility and choice. When alternatives divide neatly into (determinately) permissible and impermissible subsets, the choice prescriptions are obvious – choose a permissible alternative, refrain from the impermissible ones. But when some alternatives fall into this third category of indeterminately permissible – especially when no alternative is determinately permissible – it is much less obvious how the agent ought to choose. While a dichotomous permissibility/impermissibility partition gives rise to choice prescriptions straightforwardly, the trichotomous partition in indeterminate cases does not.

Norms for choice under indeterminacy attempt to answer this decision-theoretic challenge and plot the relationship between indeterminate permissibility and choice. Importantly, answering this question of choice requires invoking a higher-order type of normativity distinct from the lower-order norms that are subject to indeterminacy (the norms whose verdicts vary among sharpenings). While the lower-order norms are concerned with what one ought to do when all relevant facts are determinate, the higher-order norms are concerned with what one ought to do when it is indeterminate what one ought to do in the lower-order sense.[14] Let us use the term *choiceworthiness* to describe this notion of higher-order permissibility – permissibility given indeterminacy.[15]

Norms for choice under indeterminacy may be divided into two broad classes which I term *Weak norms* and *Strong norms*. Weak norms give rise to a notion of choiceworthiness that largely preserves the above trichotomy between determinately permissible, determinately impermissible, and indeterminately permissible alternatives. In particular, on weak norms all indeterminately permissible alternatives have the same normative status. In contrast, strong norms do distinguish between different indeterminately permissible alternatives and pick out a proper subset of them as more choiceworthy than the rest.

Proponents of weak norms for choice under indeterminacy include Williams (2014) and Rinard (2015). Williams (2014) puts forward a decision theory with the following consequences: (a) whenever there are any determinately permissible alternatives all and only such alternatives are choiceworthy, (b) determinately impermissible alternatives are never choiceworthy and (c) when there are no determinately permissible alternatives, all and only indeterminately permissible alternatives are choiceworthy. Assuming a straightforward link between

[14] This lower-order-higher-order relationship is analogous to the one that holds between first-order morality and what one meta-normatively ought to do. A similar relationship holds between rationality norms for choice among certain outcomes (e.g., the properties of choice functions) and norms for choice under non-moral uncertainty (e.g., expected utility theory).

[15] I follow Bales (2018) in using the term "choiceworthiness" this way. In much of the literature on moral uncertainty the term is used differently and has the sense of first-order moral value.

choiceworthiness and choice prescriptions, Williams' theory entails that one ought to choose a determinately permissible alternative when possible, and when there is no such alternative, one ought to choose some indeterminately permissible alternative.[16]

On Rinard's (2015) theory, choiceworthiness inherits the determinacy properties of lower-order permissibility.[17] So, an option may be determinately choiceworthy, indeterminately choiceworthy, or determinately unchoiceworthy according to its respective lower-order permissibility status. Assuming that one should always prefer determinately choiceworthy alternatives to others, and that one should never choose a determinately unchoiceworthy alternative over an indeterminately choiceworthy one, Rinard's theory gives rise to the same choice prescriptions as Williams'.[18] Importantly for the distinction made here, neither theory deems some indeterminately permissible alternatives as more choiceworthy than others. Decision theory, on these accounts, is silent about choices *among* indeterminately permissible alternatives.

In contrast, strong norms *do* differentiate between different indeterminately permissible alternatives and do not deem them all equally choiceworthy. When no alternatives are determinately permissible, strong norms will suggest ways of designating some proper subset of the indeterminately permissible alternatives as choiceworthy, while rejecting the rest.

[16] Williams (2014) requires that the agent choose among indeterminately permissible alternatives at random. This additional condition is consistent with the characterization of his decision theory as weak.

[17] While Williams addresses indeterminacy generally, Rinard is concerned with the case of imprecise credences, which she argues is best understood as credal indeterminacy – a state in which it is indeterminate what the agent's credences are. On this picture, the agent's representer – the set of probability functions that represents their credal state – is best understood as a set of sharpenings, that includes all precise credal states consistent with the facts. Rinard's treatment of imprecise credences therefore suggests a decision theory for indeterminacy in general, and it is this generalization I refer to here. See Bales (2018) for a discussion of Rinard and Williams as both proposing general decision theories for indeterminacy.

[18] If one abstracts away Williams' additional randomness condition.

Strong norms may distinguish between indeterminately permissible alternatives in two ways. First, *sharpening-based strong norms*, distinguish between indeterminately permissible alternatives by picking out certain "privileged" sharpenings and designating the alternatives deemed permissible on those sharpenings as more choiceworthy than the rest. So, given a privileged sharpening s^*, a strong norm of this type would deem the alternatives in $N(s^*)$ as more choiceworthy than other indeterminately permissible alternatives. Different sharpening-based strong norms propose different ways of picking out privileged sharpenings. Often though, such norms will pick a privileged sharpening on the grounds that it is "average", or "cautious", or of some other purportedly desirable property, relative to the set of all sharpenings.

Second, *sharpening-indifferent strong norms* do not commit to particular sharpenings, rather, they employ some broader aggregation procedure to discriminate between different indeterminately permissible alternatives. For example, a strong norm that considers the frequency alternatives occur as outputs of N, and deems choiceworthy the alternatives that are permissible on the greatest number of sharpenings, would be sharpening-indifferent.[19] In general, any strong norm that arrives at its verdict without committing to a particular sharpening, falls in this category of sharpening-indifferent strong norms.

While I am not aware of any proponents of strong norms for general choice under indeterminacy in the literature, such norms are widely advocated in the context of imprecise credences. In cases of imprecise credences, the agent's credal state is best

[19] Formally, for every alternative $a \in A$, let $|a|$ denote the number of sharpenings in which $a \in N(s)$ (more formally: $|a| = |\{s \in S | a \in N(s)\}|$). When a is determinately permissible $|a| = |S|$, when it is determinately impermissible $|a| = 0$, and when a is indeterminately permissible $0 < |a| < |S|$. So, the norm requires the agent to choose an action that maximizes $|a|$. Such a norm would only work when S is finite. To generalize this norm to the case of an infinite set of sharpenings a measure over sharpenings would have to be introduced.

represented by a set of credence functions – a *representor* – rather than a single credence function. In decision situations, the different credence functions in this set could give rise to different choice prescriptions (e.g., they could disagree about the expected value of the alternatives under consideration), and various norms have been suggested for dealing with such cases. Norms in this space too, may be divided into weak and strong categories depending on whether they discriminate among alternatives that are deemed permissible by some credence functions in the agent's representor and impermissible by others (these correspond to the indeterminately permissible alternatives in the indeterminacy context).[20]

It is natural to apply the norms suggested in this context to the general case of choice under indeterminacy for two reasons. First, credal imprecision is similar in important ways – structurally and otherwise – to indeterminacy: both cases involve multiple (and pairwise inconsistent) descriptions of normatively relevant facts, and it is somehow unsettled which description should be assumed for the purposes of applying the relevant norms. If the cases are similar in the right kinds of ways, then perhaps views about imprecise credences may be applied to indeterminacy (and vice versa). Second, on some accounts, credal imprecision simply *is* a case of indeterminacy, namely indeterminacy about credences.[21] If imprecision is a special case of indeterminacy, then we should expect a unified account for both cases.[22] Importantly, some of the strong

[20] Weak norms for choice with imprecise credences include Levi's (1974) *E-admissibility* and Weatherson's (2008) *Caprice*. Strong norms include *Maximin*, *α-maximin* (Hurwicz-style rule) and expected value maximization relative to an averaging of the credence functions in the agent's representor. These strong norms are sharpening-based, as they commit to a particular credence function (or a particular expected value). See Bradley (2017) and Weatherson (2008) for surveys of weak and strong norms in this context.

[21] As mentioned above, Rinard (2015) understands imprecise credences as a case of indeterminacy. See also Mahtani (2019) who argues that credence is a vague concept and the epistemic states modelled by imprecise credences are best understood as states involving vague credences.

[22] Much of what I say here applies to imprecise utility as well. However, I focus on credences because imprecise utilities are less amenable to strong norms. See Buchak (2022) for a survey of (mostly weak) norms for imprecise utility.

norms devised for imprecise credences may be inapplicable to other cases of indeterminacy. For example, some cases of indeterminacy may not allow the comparison of expected values across sharpenings that some strong norms for imprecise credence require. However, the general motivations for strong (and weak) norms in the context of credal imprecision – like the allusion to caution or moderation to resolve normative unsettledness – are naturally applicable to cases of indeterminacy.

The debate between proponents of strong and weak norms is extensive and far from settled, at least in the context of imprecise credence, and objections have been mounted at both types of norms. One central objection against weak norms is that they are indecisive and therefore insufficiently action guiding. Our decision norms – the thought goes – should resolve, not retain, the unsettled normative status of the alternatives. Just as norms for choice under uncertainty resolve uncertainty by, e.g., prescribing the alternatives that maximize expected value, norms for choice under indeterminacy should resolve indeterminacy in a similar fashion, and provide the agent with a way of choosing *among* the alternatives whose normative status is indeterminate. Instead of resolving indeterminacy, weak norms leave the agent with roughly as much guidance as they started with – the guidance yielded by the lower-order norms – and with all indeterminately permissible alternatives still on the table.[23]

In contrast, the central objection to strong norms is that they are *arbitrarily* decisive. Strong norms, the thought goes, resolve indeterminacy in arbitrary ways. When the facts that determine what one ought to do are indeterminate, we should not expect determinate and decisive verdicts as to what one ought to do. If such decisive verdicts are not based on facts about permissibility – and they cannot be based on such facts as

[23] See J. Williamson (2010) for this sort of argument. Another objection to weak norms is that they fare badly in cases involving sequences of choices. See Elga (2010) for such an objection, and Joyce (2010), Rinard (2015) and Moss (2015) for defences.

they are indeterminate – then they must be arbitrary. It is permissibility, not caution or moderation, that determines how one should act, and so the "cautiousness" or "averageness" of some sharpenings is normatively irrelevant. Strong norms, the objection goes, are biased toward decisiveness – even when the normatively relevant facts do not warrant it, they insist on yielding decisive prescriptions at the cost of arbitrariness. The indecisiveness of weak norms is therefore a feature, not a bug – it is a fitting reflection of indecisiveness in the normatively relevant facts.[24]

This debate between strong and weak norms can be cast in terms of a disagreement over the scope of facts that are relevant to choiceworthiness. Weak norms are motivated by a narrow view: only facts about the permissibility of an alternative are relevant to its choiceworthiness status. Therefore, on weak norms, choiceworthiness supervenes on permissibility. In contrast, strong norms are based on a broader view that takes facts beyond permissibility to be relevant to choiceworthiness – the type of facts that determine a privileged sharpening, or other facts that figure in the mechanisms employed by sharpening-indifferent strong norms. A broad view of this kind would leave room for distinguishing among determinately permissible alternatives as it rejects the supervenience of choiceworthiness on permissibility.

There are two types of possible arbitrariness objections to strong norms. First, one may argue that the *prescriptions* of strong norms are arbitrary because they distinguish between indeterminately permissible alternatives, and there is no non-arbitrary way of doing so. If choiceworthiness supervenes on permissibility, then strong norms must be arbitrary in this sense. Let us call this type of objection *prescription arbitrariness*. Second, one may argue that the *procedures* that strong norms employ are arbitrary, and therefore the prescriptions yielded by these procedures are arbitrary as well. For example, one

[24] See Levi (1990) and Joyce (2005) for this type of argument.

might argue that there is no non-arbitrary way of choosing a privileged sharpening and therefore sharpening-based norms are arbitrary. Let us call this type of objection *procedure arbitrariness*. The prescription arbitrariness objection applies equally to all strong norms, as it targets their defining feature – making distinctions among indeterminately permissible alternatives. In contrast, some strong norms may be more susceptible to procedure arbitrariness objections than others, as some procedures may involve more arbitrariness than others. For example, one might think that while the procedure of privileging a sharpening is always arbitrary, the procedures involved in sharpening-indifferent strong norms are not.

In sum, there are two broad approaches to decision making under indeterminacy – the weak norm approach that deems all indeterminately permissible alternatives equally choiceworthy, and the strong norm approach that offers ways of distinguishing among indeterminately permissible alternatives. Far from resolving the debate between these two approaches, my intention here is to present the dialectical landscape, so it can be applied to our case of moral uncertainty with indeterminate IVCs – an instance of choice under indeterminacy. I turn to this in the next section.

4. Strong and Weak Norms for Moral Uncertainty

With the landscape of choice under indeterminacy in the background, let us return to moral uncertainty. As argued above, if the prescriptions of the norms for moral uncertainty depend on intertheoretic value comparisons, and such comparisons are indeterminate, then the problem of moral uncertainty is an instance of choice under indeterminacy. To see this, let us apply the general framework from the previous section to this case.

The first step of applying the general framework to our case is to identify the set of sharpenings in the case of moral uncertainty. As intertheoretic value comparisons are indeterminate, each way of settling these comparisons gives rise to a sharpening of the indeterminate facts, and so the set of sharpenings in this case is given by the set of calibrations C that are left open by the facts. Recall that this set may be as large as the set of all calibrations consistent with the moral theories $C(\mathbb{V})$, and as small as a singleton including only one particular calibration (in which case IVCs are fully determinate).

Second, to make the distinction between the three statuses of lower-order permissibility – determinately permissible, determinately impermissible, and indeterminately permissible – a lower-order norm must be designated. That is, a norm that applies to cases with no indeterminacy, and can therefore partition alternatives into permissible and impermissible subsets given each calibration. For example, some philosophers think that when IVCs are determinate, agents should maximize expected value. But the framework may accommodate any lower-order norm that designates the permissible alternatives given a calibration. In general, and following the framework presented in the previous section, we can think of any such norm as function N from the set of calibrations C to subsets of A, assigning each calibration to the set of alternatives that are permissible given that calibration. Given a set of calibrations C, and a first order norm N we can then characterize the trichotomy of permissibility statuses:

1. **Determinately permissible** alternatives are those that are permissible on all calibrations: $a \in A$ is determinately permissible if for all $[\boldsymbol{v}] \in C$, $a \in N([\boldsymbol{v}])$.
2. **Determinately impermissible** alternatives are those that are impermissible on all calibrations: $a \in A$ is determinately impermissible if for all $[\boldsymbol{v}] \in C$, $a \notin N([\boldsymbol{v}])$.
3. **Indeterminately permissible** alternatives are those that are permissible on some calibrations and impermissible on others: $a \in A$ is indeterminately impermissible if there exist $[\boldsymbol{v}], [\boldsymbol{v}'] \in C$, such that $a \in N([\boldsymbol{v}])$ and $a \notin N([\boldsymbol{v}'])$.

With the problem of moral uncertainty situated in this framework of choice under indeterminacy, we can divide potential norms for moral uncertainty into weak and strong camps. A William's-style weak norm would deem all alternatives within each of the three classes above equally choiceworthy, where determinately permissible alternatives are more choiceworthy than indeterminately permissible alternatives which are more choiceworthy than determinately impermissible ones. On a Rinard-style weak norm the three classes above are also preserved, but choiceworthiness inherits the determinacy status of permissibility. On both types of norms though, when there are no determinately permissible alternatives, the agent ought to choose any indeterminately permissible alternative. So, when no alternative is determinately permissible, these norms both boil down to the following decision rule:

> **WEAK**: Choose an alternative $a \in A$ such that for some calibration $[\boldsymbol{v}] \in C$, $a \in N([\boldsymbol{v}])$.

Strong norms on the other hand, *will* make choiceworthiness distinctions among indeterminately permissible alternatives. In particular, when no alternative is determinately permissible, strong norms will prescribe choosing from some proper subset of the indeterminately permissible alternatives. *Calibration-based* strong norms (the moral uncertainty analogue of *sharpening-based*) will arrive at this subset by designating some privileged calibration $[\boldsymbol{v}^*] \in C$ and prescribing the alternatives that are permissible given that calibration. Different calibration-based strong norms will designate the privileged calibration in different ways. In general form, all such norms may be characterized relative to a privileged calibration $[\boldsymbol{v}^*]$ as follows:

> **STRONG $[\boldsymbol{v}^*]$-BASED**: choose an alternative $a \in A$ such that $a \in N([\boldsymbol{v}^*])$.

In contrast, *calibration-indifferent* strong norms (the moral uncertainty analogue of *sharpening-indifferent*) will not rely on choosing a privileged calibration. Rather, such norms would make distinctions among indeterminately permissible alternatives using other mechanisms.

How do the norms for moral uncertainty suggested in the literature fit into these categories? Surprisingly, while almost all norms for moral uncertainty with IVC indeterminacy discussed in the literature are strong, weak norms are largely absent from the debate. To demonstrate this, let me briefly survey this literature.

The strong norms suggested in the literature include both calibration-based and calibration-indifferent norms. Calibration-based norms – norms that fall into the **STRONG $[\boldsymbol{v}^*]$-BASED** schema above – pick out a particular privileged calibration and prescribe the alternatives that are permissible given that calibration. Proponents of such norms include Lockhart (2000a) who argues that moral theories should be calibrated such that the value difference between the best and worst alternatives in a given choice situation is rendered equal for all theories. Sepielli (2013) considers, without endorsing, a norm that requires calibrating theories by rendering equal the value differences between the best and worst alternatives conceivable. Most recently, Cotton-Barratt et al. (2020) argue that theories should be calibrated such that their variance is rendered equal. Other strong norms based on calibration-picking are discussed, and many others are conceptually possible.[25]

These suggestions all take the privileged calibration to embody a compromise among the moral theories the agent considers possible: the choice of the privileged calibration over all other calibrations is based on different senses of treating the moral theories "fairly" or "equally" or giving them "equal say". Indeed, all three suggestions

[25] See Sepielli (2013) and Cotton-Barratt et al. (2020) for a discussion of other calibration-based strong norms. The privileged calibrations characterized by these strong norms may all be reached by applying the following algorithm: (1) choose an arbitrary value tuple $\boldsymbol{v} = (v_1, \dots, v_n)$, (2) divide each v_i in v by some statistic s_i of v_i: its range, its standard deviation, etc. (3) the yielded tuple $\boldsymbol{v}' = (v'_1, \dots, v'_n)$, where $v'_i = v_i/s_i$ designates the privileged calibration $[\boldsymbol{v}']$. Notice that these accounts only work if the privileged calibration is a member of C. This condition is satisfied in these authors' accounts as they implicitly assume that all calibrations are left open by the facts, i.e., that $C = C(\mathbb{V})$.

favour calibrations that equalize some measure among the moral theories. The controversy between them concerns which measure's equalization best reflects treating the theories equally in the normatively relevant sense.[26]

Some philosophers worry that calibration-based norms settle IVCs arbitrarily. Much like in the general debate over norms for choice under indeterminacy, the worry is that there is no non-arbitrary way of settling indeterminate facts, and that treating moral theories equally in one sense or another, is just as arbitrary as any other way of settling indeterminate IVCs. This arbitrariness worry motivates some philosophers to adopt calibration-indifferent strong norms, that do not commit to any particular calibration and are therefore insusceptible to the arbitrariness worries about such commitments.

Calibration-indifferent strong norms make prescriptions without relying on any particular way of settling IVCs. Examples of such norms include *My Favourite Theory* (Gracely 1996, Gustafsson and Torpman 2014, Gustafsson 2022) – a norm requiring the agent to abide by the moral theory they consider most likely to be true. *My Favourite Option*[27] is another calibration-indifferent strong norm that requires the agent to choose the alternative that has the highest likelihood of being permissible (given the agent's credence). Finally, Greaves and Cotton-Barratt (2023) have recently suggested the *Nash Bargaining Norm* that requires the agent to choose an alternative that maximizes a quantity termed the *Nash Product*. While calculating the Nash product requires choosing a value tuple, the norm is nonetheless calibration-indifferent, as the choice of value tuple is inconsequential – the alternatives that maximize the Nash product are identical on any

[26] To borrow terminology from the literature on egalitarianism, proponents of these strong norms are all egalitarians when it comes to the moral theories the agent considers possible, i.e., they all agree that the privileged calibration should be one that treats all moral theories equally. Their disagreement concerns the "currency of equality" (Cohen 1989) when it comes to moral theories, i.e., the measure that should be equalized.

[27] This norm is discussed, but not endorsed, by Lockhart (2000a), Gustafsson and Torpman (2014) and Gustafsson (2022a).

value tuple used to calculate it.[28] These norms are all strong as they make distinctions among indeterminately permissible alternatives and prescribe a proper subset of them when no alternative is determinately permissible.[29] However, the prescriptions of these norms do not rely on settling IVCs in a particular way, and they are therefore calibration-indifferent.

Finally, some philosophers take the arbitrariness worries that calibration-based norms suffer from to motivate general scepticism about the metanormative project as a whole. Notably, Hedden (2016a) takes the inability to settle IVCs non-arbitrarily to suggest that the metanormative "ought" – the sense of "ought" sought after by norms for moral uncertainty – "has no important role to play in our normative theorizing",[30] and therefore the question at the centre of the metanormative project is normatively insignificant and should be abandoned.[31]

In sum, while the moral uncertainty literature is dominated by strong norms, weak norms are largely absent from the debate.[32] In contrast, as described in the previous

[28] I present and evaluate this norm in more detail in section 7.

[29] The set of indeterminately permissible alternatives is a function of the lower-order norms that govern moral uncertainty with no indeterminacy – the norms represented by N in the above definition. And since proponents of calibration-indifferent strong norms do not have to be committed to the content of N, it is not clear whether *they* would perceive their norms as strong, i.e., as making distinctions among indeterminately permissible alternatives. This is especially true of Gustafsson and Torpman (2014) who take IVCs to be always indeterminate and are thus sceptical of any norm for moral uncertainty with determinate IVCs. However, all three norms discussed in the text will tend to prescribe a very limited set of alternatives – often a set including just one alternative – which would be a proper subset of the indeterminately permissible alternatives relative to virtually any way of filling the details about N. They therefore naturally fall in this category of strong norms.

[30] Hedden (2016a), p. 104.

[31] Weatherson (2019b) argues for a similar thesis but does not take IVC indeterminacy to be supportive of it.

[32] Riedener (2021a) accommodates IVC indeterminacy when he allows for incompleteness in his "m-value" relation (the "better-than-given-the-agent's-moral-uncertainty" relation). However, when IVCs are fully indeterminate the resulting norm is silent rather than weak – it is only opinionated about pairs of alternatives about which all theories agree which is (weakly) morally better. That is, it is only opinionated when it comes to dominance relations and has no choice prescriptions when it comes to alternatives that do not stand in such relations.

section, weak norms figure quite prominently in the general literature on choice under indeterminacy. This dissimilarity is surprising, as moral uncertainty with IVC indeterminacy is an instance of choice under indeterminacy, and thus it would be reasonable to expect the dialectical landscape of the former to echo that of the latter. In particular, proponents of weak norms in the general case, should – at least *prima fascia* – endorse them in its various instances, *inter alia*, in our case of moral uncertainty. Unless IVC indeterminacy is somehow fundamentally different from other cases of indeterminacy – and I am unaware of arguments to this effect – it is hard to see how to justify the discrepancy between the two dialectical landscapes. The moral uncertainty discussion therefore appears to be skewed toward strong norms, whereas weak norms are overlooked. My purpose here is not to argue for weak norms, but to claim that they should not be overlooked and ought to be considered much more seriously than they have been thus far. In the following section I begin to fill this lacuna and examine a weak norm for the expectational case.

5. Weak Expected Value Maximization

The characterization of strong and weak norms in the previous sections has been general, placing minimal constraints on the moral theories and leaving open the content of N – the lower-order norm for moral uncertainty with no indeterminacy. In this section I zoom in to consider a weak norm for a more particular case involving additional assumptions about the structure of the moral theories, the content of the lower-order norm, and the degree of IVC indeterminacy. I will consider a weak norm for the *expectational case* – the case in which both the moral theories the agent considers possible and the lower-order norm for moral uncertainty with no indeterminacy are expectational, and IVCs are fully indeterminate.

To set up the expectational case we first need to augment the set of alternatives A to include all *probabilistic mixtures* of its elements – all lotteries that yield the alternatives in A with various probabilities. It is now assumed that the agent can choose such mixtures as well as the original "pure" alternatives in A. Formally, a probabilistic mixture is characterised by an m-tuple of non-negative numbers that sum to one: $q = (q_1, \dots, q_m)$ where q_i is the probability of alternative a_i. Let us denote the set of all such mixtures $\Delta(A)$.[33]

The domain of the moral theories is also expanded in the expectational case to include all alternatives in $\Delta(A)$, and it is assumed that all theories evaluate alternatives *expectationally* – the value of an alternative is always equal to its expected value. More formally, a value function v is *expectational* if for all $q \in \Delta(A)$, $v(q) = \sum_{i=1}^{m} q_i v(a_i)$, and a moral theory is expectational if all its elements are expectational.[34] So, in the expectational case all moral theories the agent considers possible are expectational, i.e., each $V \in \mathbb{V}$ satisfies the following condition (in addition to TRANSFORMATION and CLOSURE stated in section 2):

> EXPECTED VALUE: for any value function $v \in V$ and any mixture $q = (q_1, \dots, q_m) \in \Delta(A)$, $v(q) = \sum_{i=1}^{m} q_i v(a_i)$.

Though the assumption that the moral theories are expectational is widely adopted in the literature,[35] it is nonetheless constraining. It is not obvious that morality is risk neutral towards moral value, and opposing views may take other features of a mixture, beyond

[33] Mixtures that give probability 1 to one alternative and 0 to the rest represent "pure" alternatives – the members of A. From here onward I will use the terms *alternative* and *mixture* to refer to any element of $\Delta(A)$, the term *pure alternative* to refer to the original alternatives, and the term *proper mixture* to refer mixtures that give positive probability to more than one alternative.

[34] If a value function is expectational then so are all its positive affine transformations. So, if any member v of a moral theory V is expectational, all members are.

[35] See Ross (2006), Cotton-Barratt et al. (2020), MacAskill et al. (2020), Riedener (2020, 2021), Greaves and Cotton-Barratt (2023).

its expectation, to be morally relevant. For example, morality might be risk-averse toward moral value, such that the "spread" or "riskiness" of a probabilistic mixture detracts from its moral value (*ceteris paribus*).[36] However, much of the discussion of cardinal moral uncertainty incorporates this assumption and so I will do so here as well.

The second assumption that characterizes the expectational case, is that N – the norm governing moral uncertainty with no indeterminacy – is expectational. That is, when IVCs are determinate and therefore characterized by a unique calibration, the agent is required to choose an alternative with maximal expected value on that calibration. In other words, given a calibration, the agent is required to choose an alternative that is *maximal on that calibration*:

> MAXIMALITY ON A CALIBRATION: an alternative $q \in \Delta(A)$ is *maximal on calibration* $[\boldsymbol{v}] \in C(\mathbb{V})$ iff for some $(v_1, \dots, v_n) \in [\boldsymbol{v}]$, all alternatives $q' \in \Delta(A)$ are such that $\sum_{i=1}^{n} p_i v_i(q) \geq \sum_{i=1}^{n} p_i v_i(q')$.[37]

Let E be a function from calibrations to sets of alternatives, that maps each calibration to the set of alternatives that is maximal on that calibration. With these definitions in place, we may formulate the following norm for moral uncertainty when IVCs are determinate, and thus specified by a single calibration:

> EXPECTED VALUE MAXIMIZATION (EVM): given a specified calibration $[\boldsymbol{v}] \in C(\mathbb{V})$, the agent ought to choose an alternative in $E([\boldsymbol{v}])$.

This second expectational assumption is also controversial. As mentioned above, some philosophers argue that IVCs are never determinate, and therefore EVM is never

[36] For decision theories that relax this assumption of risk neutrality see Buchak (2013), Stefánsson & Bradley (2015), and Goldschmidt and Nissan-Rozen (2021).

[37] Notice that if a mixture has maximal credence-weighted expected value on one element of a calibration, then this holds for all elements, because the order of expected values is preserved under positive affine transformation. For this reason, maximality is defined relative to a calibration rather than a value tuple.

applicable even if it is correct. Further, the same worries about attitudes to risk that challenge the assumption that moral theories are expectational, apply here as well. One could argue that risk-neutrality toward intertheoretic moral value is not warranted, and that metanormativity should allow risk aversion or sensitivity to non-expectational features of a mixture more generally.[38] Nonetheless, this assumption is accepted throughout much of the discussion of moral uncertainty with indeterminate IVCs and is presupposed by proponents of most strong norms.[39]

Finally, I will assume that IVCs are fully indeterminate – all calibrations consistent with the moral theories are left open by the facts and thus $C = C(\mathbb{V})$. This assumption obviously constrains the case under consideration but is also quite commonly adopted in the literature on IVC indeterminacy.

With the expectational case defined, we may partition alternatives into the three permissibility classes discussed above, relative to the set of indeterminate calibrations $C(\mathbb{V})$:

1. **Determinately permissible** alternatives are those that are maximal on all calibrations: $q \in \Delta(A)$ is determinately permissible if for all $[\boldsymbol{v}] \in C(\mathbb{V})$, $q \in E([\boldsymbol{v}])$.
2. **Determinately impermissible** alternatives are those that are maximal on no calibrations: $q \in \Delta(A)$ is determinately permissible if for all $[\boldsymbol{v}] \in C(\mathbb{V})$, $q \notin E([\boldsymbol{v}])$.
3. **Indeterminately permissible** alternatives are those that are maximal on some calibrations and not maximal on others: $q \in \Delta(A)$ is indeterminately impermissible if there exist $[\boldsymbol{v}], [\boldsymbol{v}'] \in C(\mathbb{V})$, $q \in E([\boldsymbol{v}])$ and $q \notin E([\boldsymbol{v}'])$.

[38] See Nissan-Rozen (2015a) for an argument against EVM that stems from this point about risk attitudes.

[39] In particular, all proponents of calibration-based strong norms mentioned above – Lockhart (2000a), Sepielli (2013), Cotton-Barratt et al. (2020) – assume EVM.

In the interesting cases, in which there are no determinately permissible alternatives, the property of being indeterminately permissible is identical to *maximizability*:

MAXIMIZABILITY: a mixture is *maximizable* iff it is maximal on *some* calibration.[40]

Now, assuming there are no determinately permissible alternatives, we may state the expectational variant of **WEAK** from the previous section, by situating E in place of N:

WEAK EVM: Choose an alternative $q \in \Delta(A)$ such that for some calibration $[\boldsymbol{v}] \in C(\mathbb{V})$, $q \in E([\boldsymbol{v}])$. Or, more succinctly: choose a maximizable alternative.

Like weak norms for moral uncertainty in general, WEAK EVM is not argued for in the literature. Instead, philosophers tackling the expectational case argue for various strong norms (more on this in the next section). However, as argued for in the previous section, weak norms deserve to be taken more seriously in this debate, and proponents of weak norms for general indeterminacy, it seems, should endorse WEAK EVM in the expectational case. A proponent of weak norms might argue for WEAK EVM by arguing that maximizability is necessary and sufficient for choiceworthiness. First, an argument for necessity:

(1) If an alternative is not maximizable then it is determinately impermissible.
(2) If an alternative is determinately impermissible then it is not choiceworthy.
(3) Therefore, maximizability is necessary for choiceworthiness.

Second, an argument for sufficiency:

(1) The choiceworthiness status of an alternative supervenes on its permissibility status.

[40] Equivalently, in terms of value tuples, a mixture $q \in \Delta(A)$ is *maximizable* iff there exists a value tuple $\boldsymbol{v} = (v_1, \dots, v_n) \in T(\mathbb{V})$, such that for all alternatives $q' \in \Delta(A)$, $\sum_{i=1}^{n} p_i v_i(q) \geq \sum_{i=1}^{n} p_i v_i(q')$.

(2) Therefore, if some maximizable alternative is choiceworthy, they all are.

(3) Some alternatives are choiceworthy.

(4) If an alternative is choiceworthy then it is maximizable (from necessity argument).

(5) Therefore, some maximizable alternative is choiceworthy (from 3 and 4).

(6) Therefore, Maximizability is sufficient for choiceworthiness (from 2 and 5).

Let us call these arguments the *standard arguments for necessity and sufficiency* respectively. Proponents of strong norms would likely accept necessity and reject sufficiency. In particular, they would reject the supervenience thesis in the first premise of the argument for sufficiency. My purpose in presenting these arguments is not to argue for WEAK EVM or take a stand on the general divide between strong and weak norms, but to encourage a debate that takes this norm seriously. My goal is to provide the infrastructure for a moral uncertainty debate that is more informed by, and more similar to, the general debate about choice under indeterminacy.

However, examining WEAK EVM more closely reveals that it possesses an important property that allows the formulation of a stronger argument for necessity than the standard argument above. In particular, WEAK EVM is equivalent to the norm prohibiting the choice of *dominated* alternatives. To see this, let us define the following notion of *dominance*:

> DOMINANCE: for any two mixtures $q, q' \in \Delta(A)$, q *dominates* q' iff for all value tuples $\boldsymbol{v} = (v_1, \ldots, v_n) \in T(\mathbb{V})$ the following two conditions hold:
>
> (1) $v_i(q) \geq v_i(q')$ for all $i \in \{1, \ldots, n\}$.
>
> (2) $v_j(q) > v_j(q')$ for some $j \in \{1, \ldots, n\}$.
>
> An alternative is *dominated* if there exists some other available alternative that dominates it.

Now, the following theorem ensures the stated equivalence:

Theorem 2.1: an alternative is maximizable if and only if it is not dominated.[41]

This Theorem has two upshots. First, it gives rise to an argument for maximizability being necessary for choiceworthiness, and thus in favour of WEAK EVM as a necessary condition on choice. Second, the theorem helps illuminate the space of strong norms for the expectational case. I devote the remainder of this section to discuss the first upshot, and the next section to discuss the second upshot.

The theorem gives rise to the following alternative argument for the necessity of maximizability for choiceworthiness:

(1) If an alternative is not maximizable then it is dominated.
(2) If an alternative is dominated, then it is not choiceworthy.
(3) Therefore, maximizability is necessary for choiceworthiness.

Let us call this argument the *dominance-based argument for necessity*. Importantly, this argument does not rely on any assumption about the lower-order norms governing moral uncertainty with no indeterminacy – i.e., assumptions regarding the content of N. If successful, the dominance-based argument suggests that *regardless of the content of the lower-order norms*, agents should maximize expected value according to some calibration. That is, we do not have to agree on whether EVM is true to agree that WEAK EVM is a necessary condition on choice, and therefore that the agent should always choose a maximizable alternative.

In contrast, the *standard argument for necessity* relies heavily on EVM – on the assumption that when IVCs are determinate, agents ought to maximize expected value. The first premise of the argument – the claim that non-maximizable alternatives are determinately impermissible – is only plausible if EVM is true. The dominance-based

[41] See appendix A (A1) for proof.

argument replaces the EVM assumption in the standard argument with a premise linking domination and choiceworthiness (premise 2). Therefore, the strength of the dominance-based argument relative to the standard argument will depend on the plausibility of the domination-choiceworthiness link in premise 2 relative to EVM. At the very least, the dominance-based argument provides a path to WEAK EVM for EVM sceptics.

Let us consider the dominance-based argument more closely. Are the premises plausible? We need not worry about the first premise, as it is guaranteed by theorem 2.1, but the second premise – which is doing the normative heavy-lifting – must be justified for the argument to be forceful. The prohibition of choosing dominated alternatives is quite plausible and fairly broadly accepted.[42] Importantly, its plausibility is entirely independent of whether EVM is the correct norm for cases with no indeterminacy. To see this, I offer two brief arguments for the claim that agents should never choose dominated alternatives.

First, that the agent ought to refrain from choosing dominated alternatives follows from the following two principles:

> CERTAIN PERMISSIBILITY (CP): When choosing between two actions – ϕ and ψ – if the agent is certain that ϕ-ing is permissible but is uncertain whether ψ-ing is permissible, then they ought to ϕ.
>
> OUGHT CONSISTENCY (OC): If the agent ought to choose x when choosing among two options x and y, then they ought not choose y whenever x is available.

CP is a principle constraining the relation between an agent's first-order beliefs about permissibility and what they ought to do (metanormatively) given those beliefs. While it is not uncontroversial, it is a very weak constraint, as it only requires avoiding

[42] See MacAskill (2013) and Hicks (2018). Hedden (2016a) accepts a broad scope version of dominance.

the risk of impermissibility when this is possible. Any view that takes the agent's moral beliefs to constrain what they ought to do in *some* way, would likely accept this constraint.

OC is a consistency condition on the metanormative ought – the type of "ought" relevant to moral uncertainty. The principle ensures that whether the agent ought to choose x over y is independent of the availability of other options. It is a metanormative version of a common constraint on rational choice.[43]

To see how premise 2 follows from these two principles, let us suppose q dominates q' and the agent must choose between them. The agent is certain that choosing q over q' is permissible because they are certain that q is morally better than, or as equally good as q', and in both cases this would render choosing q permissible. However, the agent is not certain that q' is permissible since they give positive credence to the proposition that q' is morally worse than q (in which case it would not be permissible to choose q'). Therefore, according to CP, the agent ought to choose q over q'. It then follows from OC that the agent ought not to choose q' whenever q is available. Since an option is only dominated if there exists an available option that dominates it, it follows that the agent should never choose a dominated option.

The second argument for premise 2 relies on the agent's subjective reasons for choosing a dominated option.[44] Suppose again that q dominates q' and the agent must choose between them. In such a case the agent appears to possess two types of reasons relevant to the choice at hand – reasons for having credence in the proposition "q and q' are equally good morally" and reasons for having credence in the proposition "q is morally better than q'". The first type of reasons are reasons for choosing either option

[43] OC follows from property α of choice functions discussed in Sen (1993) (see references there).

[44] I use the term "subjective reason" in Schroeder's (2008) sense: a subjective reason is a proposition that the agent believes, and if it is true, it is an objective reason. MacAskill (2013) frames a dominance principle in terms of subjective reasons.

and the second type of reasons are reasons for choosing q over q'. Therefore, every reason the agent has for choosing q' is also a reason for choosing q, but there are additional reasons for choosing q that are not reasons for choosing q'. Therefore, the agent has reasons for choosing q over q' but no reasons for choosing q' over q, and thus they should choose q. Together with OC, this line of reasoning also gives rise to premise one.

So, refraining from choosing dominated alternatives is a plausible requirement on any view that takes the agent's moral beliefs, or their subjective reasons for those beliefs, to constrain what they ought to do. Therefore, on any such view the dominance-based argument should go through and its conclusion – WEAK EVM as a necessary condition on choice – should be accepted. Importantly, these arguments only entail that maximizability is necessary for choiceworthiness, not that it is sufficient. Proponents of strong norms may still hold that only some proper subset of the maximizable alternatives is choiceworthy. However, the argument does entail the rejection of strong norms that prescribe non-maximizable alternatives. In other words, a consequence of the argument is that agents ought to choose as if they are maximizing expected value on some calibration. And this conclusion does not rely on any assumption about what agents ought to do when IVCs are determinate.

In sum, WEAK EVM should be considered more seriously as a norm for the expectational case, and the standard arguments for the necessity and sufficiency of maximizability for choiceworthiness presented above are good starting points for such a consideration. Theorem 2.1 shows that WEAK EVM is equivalent to the norm prohibiting the choice of dominated alternatives. This equivalence gives rise to the dominance-based argument for WEAK EVM as a necessary condition for choice in the expectational case. This argument entails that agents should maximize expected value on some calibration without assuming anything about the norms governing moral uncertainty with determinate IVCs.

6. The Space of Strong Norms

The second upshot of theorem 2.1 is to illuminate the relationships between different types of strong norms for the expectational case. Recall that strong norms come in two flavours – calibration-based and calibration-indifferent. Calibration-based strong norms are characterized in terms of a privileged calibration $[\boldsymbol{v}^*]$ and the lower-order norm N, so in the expectational case, replacing N with E in the schema from section 4 yields the following characterization of calibration-based strong norms:

STRONG $[\boldsymbol{v}^*]$-BASED EVM: choose an alternative $q \in \Delta(A)$ such that $q \in E([\boldsymbol{v}^*])$.

In other words, calibration-based strong EVM norms require the agent to choose an alternative that is maximal on the privileged calibration. Different calibration-based norms differ in the method for choosing the privileged calibration. Indeed, the calibration-based strong norms proposed in the literature and surveyed in section 4 are all EVM norms. In contrast, calibration-indifferent strong norms are not committed to any calibration and may prescribe in principle any set of alternatives.

Theorem 2.1 sheds some light on the relation between calibration-indifferent norms that never prescribe dominated alternatives, and calibration-based norms. In particular, the theorem implies that any alternative prescribed by a calibration-indifferent strong norm – so long as it is non-dominated – is also prescribed – possibly among other alternatives – by some calibration-based strong norm. In other words, for every non-dominated alternative $q \in \Delta(A)$ that might be deemed choiceworthy by a calibration-indifferent strong norm, there is a calibration $[\boldsymbol{v}]$ such that q is maximal on $[\boldsymbol{v}]$, and would therefore be deemed choiceworthy by the strong $[\boldsymbol{v}]$-based EVM norm.

Importantly, this result does not imply that any calibration-indifferent norm that never prescribes dominated alternatives is extensionally equivalent to some calibration-based norm. This is because while calibration-indifferent norms could in principle deem any set of non-dominated alternatives as choiceworthy, calibration-based EVM norms are more constrained. For example, calibration-based EVM norms can never prescribe a proper mixture alone, without the pure alternatives that it gives positive probability to. More generally, the set of maximal elements on any calibration will always satisfy a condition I term *pure convex closure* which I will now define.

To define this condition, let me first introduce some further technical notions. Let the *pure components* of an alternative $q \in \Delta\,(A)$ be the pure alternatives that receive positive probability on q. Formally, the set of q's pure components, denoted $P(q)$, equals $\{a_i \in A | q_i > 0\}$.[45] The pure components of a set of alternatives $X \subseteq \Delta\,(A)$ are the union of the pure components of its elements: $P(X) = \bigcup_{q \in X} P(q)$. The *convex closure* of a set of pure alternatives $Y \subseteq A$ denoted $Con(Y)$ is the set of mixtures that give zero probability to all pure alternatives not in Y, formally: $Con(Y) = \{(q_1, \dots, q_m) \in \Delta(A) | a_i \notin Y \rightarrow q_i = 0\}$.[46] The *pure convex closure* of a set of (not necessarily pure) alternatives $X \subseteq \Delta\,(A)$ is the convex closure of its pure components, and is denoted: $Con\big(P(X)\big)$. A set of alternatives is *purely convexly closed* (or satisfies *pure convex closure*) if it is equal to its pure convex closure. The prescriptions of calibration-based EVM norms, that is, the images of E, satisfy the following condition:

Claim 1: For every calibration $[\boldsymbol{v}] \in C(\mathbb{V})$, $E([\boldsymbol{v}])$ is purely convexly closed.[47]

[45] For simplicity, I am assuming throughout this section that A includes no *superfluous* pure alternatives, where a pure alternative is superfluous, if all moral theories agree that its value is identical to some other pure or mixed alternative. Formally, $a \in A$ is superfluous if there exists $q \in \Delta(A)$ such that $a \neq q$ but for every value tuple $v = (v_1, \dots, v_n)$ and for every $1 \leq i \leq n$, $v_i(q) = v_i(a)$. This assumption may be dropped if the definitions in the text are changed so that pure components include only non-superfluous pure components.

[46] The definition in the text is given in terms of pure alternatives for simplicity. More generally, the convex closure of a set of (not necessarily pure) alternatives $X \subseteq \Delta\,(A)$ is defined as follows. Let $Inf_i(X)$ and $Sup_i(X)$ be the infimum and supremum (respectively) of the set of probabilities that the elements of X give to the pure alternative a_i. The convex closure of X is: $Con(X) = \{q = (q_1, \dots, q_m) \in \Delta(A) | Inf_i(X) \leq q_i \leq Sup_i(X)\}$. When $X \subseteq A$, this definition is equivalent to the one in the text.

[47] See appendix A (A2) for proof.

Claim 1 entails that if a calibration-indifferent strong norm I deems a set of alternatives choiceworthy, and that set is not purely convexly closed, then no calibration-based EVM norm would prescribe that precise set, and therefore, no calibration-based EVM norm would be extensionally equivalent to I. However, I's prescription may still be *included* among the alternatives deemed choiceworthy by some calibration-based EVM norm. This will be the case if and only if I's output – i.e., the set of alternatives it deems choiceworthy – is such that its pure convex closure in non-dominated:[48]

> **Claim 2**: let $X \subseteq \Delta(A)$ be some nonempty set of alternatives. The pure convex closure of X is non-dominated if and only if there exists a calibration $[\boldsymbol{v}]$ such that $X \subseteq E([\boldsymbol{v}])$.[49]

If, on the other hand, I's output is non-dominated, but violates the condition in claim 2, then while theorem 2.1 ensures that all of its prescribed alternatives are maximizable, they will not be maximal on any single calibration. So, we may characterize three types of relations that dominance-respecting calibration-indifferent norms may bear towards calibration-based norms. Let X be the set of alternatives prescribed by a dominance-respecting calibration-indifferent norm I, then:

1. **Extensional equivalence**: I is *extensionally equivalent* to some calibration-based norm if its prescriptions are identical to the prescriptions of some calibration-based norm. That is, if there exists a calibration $[\boldsymbol{v}]$ such that $X = E([\boldsymbol{v}])$.
2. **Subsumption**: I is *subsumed* by some calibration-based norm if its prescriptions are included in the prescriptions of some calibration-based norm. That is, if there exists a calibration $[\boldsymbol{v}]$ such that $X \subseteq E([\boldsymbol{v}])$.

[48] A set of alternatives is non-dominated if all of its elements are non-dominated.

[49] See appendix A (A3) for proof.

3. **Joint Subsumption:** I is *jointly subsumed* by some calibration-based norms if its prescriptions are included in the prescriptions of some, possibly multiple, calibration-based norms. That is, if there exists a set of calibrations Y such that $X \subseteq \bigcup_{[\boldsymbol{v}] \in Y} E([\boldsymbol{v}])$.

The three relations are listed in a descending order of strength: *Extensional equivalence* entails *Subsumption* which entails *Joint subsumption*. All dominance-respecting strong norms satisfy *Joint Subsumption*, and some such norms satisfy the stronger conditions. In particular, a dominance-respecting calibration-indifferent strong norm satisfies *Subsumption* if and only if, the pure convex closure of its prescriptions is non-dominated (claim 2).[50] I leave the condition separating *Subsumption* from *Extensional equivalence* for future work, but I conjecture that I (a dominance-respecting norm) is *extensionally equivalent* to some calibration-based norm if and only if, X is purely convexly closed.[51]

7. The Bargaining Norm

The general insights about the relationship between the two types of strong norms are helpful in evaluating a strong norm recently suggested by Hilary Greaves and Owen Cotton-Barratt (2023). These authors suggest applying the Nash (1950) bargaining model – originally devised to represent and analyse a bargaining situation between two agents – to the case of moral uncertainty.[52] This model, applied to the expectational case characterized above, gives rise to a calibration-indifferent strong norm that I will term the

[50] Norms that prescribe a single alternative always satisfy *Subsumption*, because if X is a singleton (and non-dominated) then its pure convex closure is non-dominated, and therefore *Subsumption* applies.

[51] Formally, the conjecture to be determined is: "a set of alternatives $X \subseteq \Delta(A)$ is non-dominated and purely convexly closed if and only if there exists a calibration $[\boldsymbol{v}]$ such that $X = E([\boldsymbol{v}])$".

[52] More precisely, the authors apply the a-symmetric version of the model suggested by Harsanyi and Selten (1972). See also references in Greaves and Cotton-Barratt (2023), p. 15.

Bargaining norm. To state the Bargaining norm, in addition to the formal apparatus presented above, the notion of a *disagreement point* must be introduced. A disagreement point d is an n-tuple of alternatives that is strictly dominated by some alternative: $d = (d_1, \dots, d_n)$ where $d_i \in \Delta(A)$ for $i = 1, \dots, n$, and there exists an alternative $q \in \Delta(A)$ such that $v_i(q) > v_i(d_i)$ for $i = 1, \dots, n$. Given a choice of a disagreement point d, the Bargaining norm requires choosing an alternative q that maximizes the Nash Product:

$$\prod_{i=1}^{n} \big(v_i(q) - v_i(d_i)\big)^{p_i}$$

Where the v_is are given by an arbitrary value tuple. The Bargaining norm has several desirable features. First, it is a calibration-indifferent norm – even though the calculation of the Nash product requires choosing a value tuple, this choice is inconsequential as the alternatives that maximize the Nash product are the same on all value tuples. So, in contrast to calibration-based EVM norms, the Bargaining norm is not committed to any particular calibration and is therefore immune to criticisms about settling inter-theoretic value comparisons arbitrarily. Second, the Bargaining norm never prescribes dominated alternatives, as the Nash product of such an alternative will never be maximal – it will always be smaller than that of the alternatives that dominate it.[53]

To see how the Bargaining norm relates to calibration-based EVM norms, consider the following example. Assume the agent divides their credence equally among only two moral theories that evaluate six alternatives, so $\mathbb{V} = \{V_1, V_2\}$, $p = (0.5, 0.5)$ and $A = \{a_1, \dots, a_6\}$. And let $\boldsymbol{v} = (v_1, v_2)$ be some value tuple of the moral theories that evaluates the alternatives as follows:

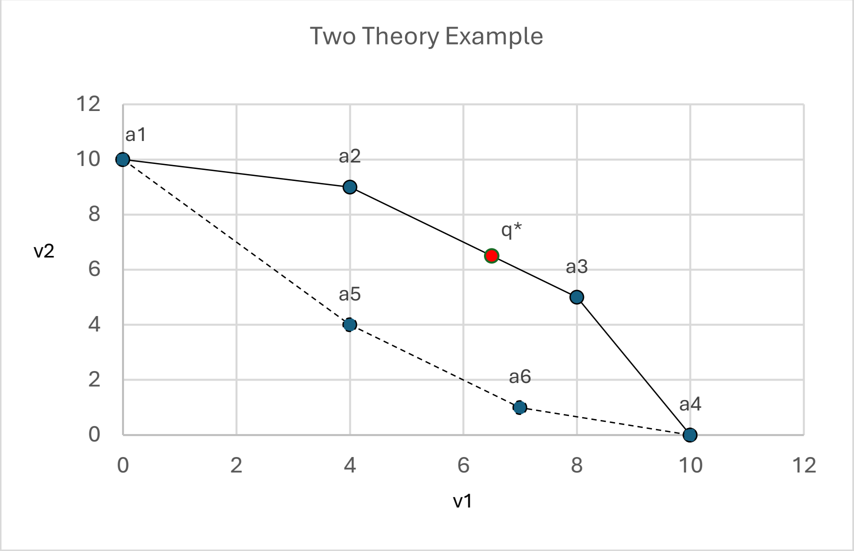
[53] For other desirable features of the Nash norm see Greaves and Cotton-Barratt (2023).

TABLE 2.3: *Two-theory example*

	v_1	v_2
a_1	0	10
a_2	4	9
a_3	8	5
a_4	10	0
a_5	5	5
a_6	7	2

Since our example involves only two moral theories, it lends itself well to graphic presentation. The graph below plots the values of each alternative according to the two theories, where the x axis represents v_1 and the y axis represents v_2. The points (except q^*) represent the values of the pure alternatives, and the hexagon created by them is the space of values of all mixtures involving them (the space of values of $\Delta(A)$). Finally, the solid line, sometimes termed the *Pareto frontier*, includes the values of all non-dominated alternatives. Before applying the Bargaining norm to this example, let me briefly introduce a simpler notation for proper mixtures that ignores the alternatives that receive probability zero, and lists the positive-probability alternatives with their probabilities. So, for example the mixture that gives probability 0.3 to a_2 and 0.7 to a_6, will be denoted $(0.3a_2, 0.7a_6)$ instead of $(0,0.3,0,0,0,0.7)$.

FIGURE 2.1: *Two-theory example*



To apply the Bargaining norm to this case, we must first designate a disagreement point. For now, let the disagreement point be $d = (a_1, a_4)$, and so the value vector of d is (0,0). A somewhat tedious calculation then yields that the mixture that maximizes the Nash product $\left(v_1(q) - v_1(a_1)\right)^{0.5}\left(v_2(q) - v_2(a_4)\right)^{0.5}$ is $q^* = \left(\frac{3}{8}a_2, \frac{5}{8}a_3\right)$, which gives positive probability to only a_2 and a_3, and therefore lies between them in the graph above.

How does this prescription relate to various calibration-based norms? Theorem 2.1 entails that there exists a calibration on which q^* is maximal, but since q^* is not purely convexly closed, it is never uniquely maximal. Indeed, on the calibration in the example, the pure convex closure of q^* is maximal. That is, a_2, a_3, and all mixtures involving only them – all points lying between them – have maximal expected value. Of course, the maximality of a_2, a_3, and their pure convex closure is not preserved across all calibrations. In fact, all purely convexly closed and non-dominated sets of alternatives in this example are maximal on different calibrations. There are seven such sets: $\{a_1\}$, $[a_1, a_2]$, $\{a_2\}$,

$[a_2, a_3]$, $\{a_3\}$, $[a_3, a_4]$, $\{a_4\}$.[54] To see how different calibrations deem these sets maximal, we can express calibrations in terms of the original value tuple $\boldsymbol{v}$ in the example, and a positive scaling vector $\boldsymbol{x} = (x_1, x_2)$ where $\boldsymbol{xv} = (x_1 v_1, x_2 v_2)$ is the value tuple yielded by multiplying the value function v_i by the scalar x_i. In our two-theory case, we can reach all non-dominated alternatives by scaling only one of the theories, so let us fix $x_1 = 1$ and generate calibrations by choosing different values for x_2. The following table describes how the maximal alternatives vary as a function of x_2 where $x_1 = 1$:

TABLE 2.4: *EVM Prescriptions as a function of calibrations*

Value of $\boldsymbol{x_2}$	$\boldsymbol{E}([\boldsymbol{xv}])$
$x_2 > 4$	$\{a_1\}$
$x_2 = 4$	$[a_1, a_2]$
$1 < x_2 < 4$	$\{a_2\}$
$x_2 = 1$	$[a_2, a_3]$
$0.4 < x_2 < 1$	$\{a_3\}$
$x_2 = 0.4$	$[a_3, a_4]$
$x_2 < 0.4$	$\{a_4\}$

The value of x_2, or more precisely the ratio between x_2 and x_1, can be thought of as the degree to which the moral theory V_2 is "amplified" relative to V_1.[55] Accordingly, the prescriptions of calibration-based EVM norms move along the pareto frontier as the calibration changes, such that the more V_2 is amplified – that is, the greater the ratio x_2/x_1 – the better these prescriptions are in the lights of V_2. So, from prescribing a_1 – the best

[54] I am using the notation $[a_i, a_j]$ where a_i and a_j are pure alternatives, to denote their convex closure – the set of mixtures that give positive probability to them and zero probability to the rest. Graphically, the values of the alternatives in $[a_i, a_j]$ are represented by the line connecting the value vector of a_i to the value vector of a_j.

[55] Importantly, this notion of amplification is purely relative – it is stated in terms of the original value tuple $\boldsymbol{v}$, and as inter-theoretic value comparisons are indeterminate, there is nothing special about $\boldsymbol{v}$ (or its calibration $[\boldsymbol{v}]$).

alternative according to V_2 – on calibrations that amplify V_2 the most, the prescriptions of calibration-based EVM norms gradually deteriorate (à la V_2) as V_2 becomes less amplified, until they reach a_4 – the worst alternative according to V_2. So, the prescriptions of calibration-based EVM will move across the Pareto frontier as calibrations amplify and de-amplify the different theories.

This comparison between the Bargaining norm and calibration-based EVM norms highlights the purported advantage of the calibration-indifference of the former. While the prescriptions of calibration-based norms depend so thoroughly on the choice of a calibration – indeed, prescriptions may vary across the whole Pareto frontier as a function of this choice – the verdicts of the Bargaining norm are free from any such choice. Therefore, if there is no non-arbitrary way of settling IVCs, then the calibration-based EVM norms are susceptible to an arbitrariness objection that the Bargaining norm is immune to, namely, that their prescriptions rely on arbitrarily settling IVCs in a certain way. Importantly, there is nothing wrong with norms relying on inconsequential arbitrary choices, like choosing a value tuple to calculate the Nash product. The arbitrariness objection to calibration-based norms is driven by the *combination* of arbitrariness and consequentialism involved in choosing a calibration.

Importantly, the arbitrariness objection that might motivate the adoption of the Bargaining norm over calibration-based EVM norms involves *procedure arbitrariness*, not *prescription arbitrariness*. As discussed above, all strong norms are susceptible to *prescription arbitrariness* objections as they all rely on facts beyond permissibility to distinguish among indeterminately permissible alternatives. So, if the Bargaining norm has an arbitrariness-related advantage over calibration-based norms it is only with respect to *procedure arbitrariness* objections. If the process of picking a privileged calibration is necessarily arbitrary, then it seems that calibration-based EVM norms are susceptible to an objection that the Bargaining norm is immune to. Indeed, Greaves and

Cotton-Barratt take the Bargaining norm's calibration-indifference and its evasion of the need to settle IVCs, to be a central advantage over calibration-based norms.[56]

However, while the Bargaining norm is not susceptible to the arbitrariness objection relating to calibration choice, it *is* susceptible to a different *procedure arbitrariness objection*. Recall that the Bargaining norm relies on the designation of a *disagreement point* – an additional piece of formalism, which did not appear in the initial model and is not alluded to by any of the other norms considered thus far. While in Nash's original model, the disagreement point represents the payoffs the agents receive in absence of agreement, in the context of moral uncertainty, the point has no obvious referent. In contrast to the other components of the formalism – the alternatives, the moral theories, and the credence tuple – that all represent components in the "target system" the formalism is meant to model, the disagreement point does not appear to correspond to anything of normative import. Therefore, the choice of a particular disagreement point in the context of moral uncertainty must be arbitrary. In contrast to the original bargaining context where the disagreement point would presumably be determined by some empirical fact about the consequences of a failure to reach an agreement, there is no analogous determining fact in the moral uncertainty context. And absent such a fact determining the identity of the disagreement point, any designation of it would be arbitrary.

Now, recall that arbitrariness can be innocuous. If the arbitrary choice is relatively inconsequential and does not strongly affect the outcome of the relevant process, then the arbitrariness is not objectionable. In other words, arbitrariness in a procedure is objectionable only to the extent that it is consequential. So, the mere arbitrariness of the designation of the disagreement point is not necessarily grounds for an arbitrariness *objection* to the Bargaining norm. However, to the determent of the Bargaining norm, the

[56] Greaves and Cotton-Barratt (2023), p. 6.

choice of a disagreement point is highly consequential. Varying the disagreement point can lead the Bargaining norm to prescribe almost any non-dominated alternative. To see this, let us reformulate the Bargaining norm as a function B from disagreement points to sets of alternatives that maps each disagreement point to the set of alternatives that maximize the Nash product on that disagreement point:

$$B(d) = \underset{q \in \Delta(A)}{\operatorname{argmax}} \prod_{i=1}^{n} \big(v_i(q) - v_i(d_i)\big)^{p_i}$$

To demonstrate the sensitivity of the Bargaining norm's prescriptions to the choice of the disagreement point, let us consider varying the disagreement point in the example above. Instead of $d = (a_1, a_4)$ I consider in the table below five other values for d and the prescriptions they give rise to. Since all the disagreement points considered below use the same alternative in both entries, i.e., they satisfy $d_1 = d_2$, I list them by simply stating the relevant alternative q instead of the pair (q, q):[57]

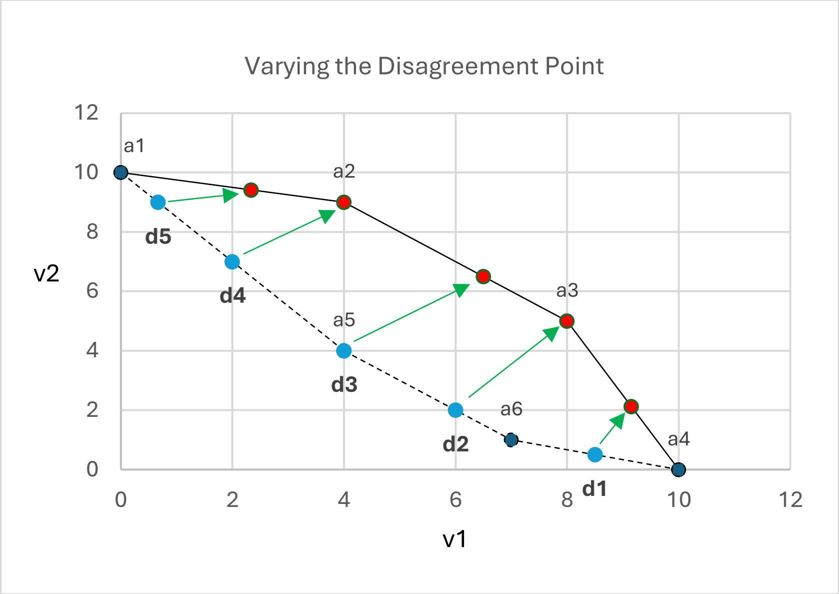
TABLE 2.5: *Bargaining prescriptions as a function of disagreement points*

Index	Designated $\boldsymbol{d}$	Value vector of $\boldsymbol{d}$	$\boldsymbol{B(d)}$
1	$\left(\frac{1}{2}a_3, \frac{1}{2}a_4\right)$	(8.5,0.5)	$(0.425a_3, 0.575a_4)$
2	$\left(\frac{1}{3}a_5, \frac{2}{3}a_6\right)$	(6,2)	a_3
3	a_5	(4,4)	$\left(\frac{3}{8}a_2, \frac{5}{8}a_3\right)$
4	$\left(\frac{1}{2}a_1, \frac{1}{2}a_5\right)$	(2,7)	a_2
5	$\left(\frac{5}{6}a_1, \frac{1}{6}a_5\right)$	$\left(\frac{2}{3}, 9\right)$	$\left(\frac{5}{12}a_1, \frac{7}{12}a_2\right)$

[57] Disagreement points that satisfy $d_1 = d_2$ (and in the general case $d_1 = \cdots = d_n$) represent alternatives. That is, for such points, there is an alternative q such that $v_i(d_i) = v_i(q)$ for $i = 1, \ldots, n$. In contrast, disagreement points that violate this condition will not correspond to any alternative. For example, the disagreement point $d = (a_1, a_4)$ stipulated above – there is no alternative q such that $v_1(q) = v_1(a_1)$ and $v_2(q) = v_2(a_4)$. Graphically, the points that satisfy the above condition will be in the hexagon created by the pure alternatives (their convex closure), and the points that violate it will be outside it.

These results are depicted in the graph below with arrows connecting each disagreement point (d1-d5) to its respective Bargaining prescription:

FIGURE 2.2: *Bargaining prescriptions as a function of disagreement points*



As illustrated by this example, the prescriptions of the Bargaining norm vary greatly with the choice of the disagreement point. Indeed, in the above example, any point on the Pareto frontier, except its extremes a_1 and a_4, may be prescribed by the Bargaining norm given some disagreement point.[58] So, the choice of the disagreement point is highly consequential, and therefore, the Bargaining norm is susceptible to a procedure arbitrariness objection, as it involves an arbitrary and consequential choice. Furthermore, choosing a disagreement point appears to be roughly *as* consequential as choosing a calibration for a calibration-based norm. The outputs of both functions E – from calibrations to maximal alternatives – and B – from disagreement points to their Bargaining prescriptions – span all, or almost all, of the Pareto frontier.

[58] However, proper mixtures that give a_1 or a_4 a probability arbitrarily close to 1 (but strictly less than 1) can still be prescribed by the Bargaining norm given the right disagreement points.

The arbitrariness and consequentialism of choosing a disagreement point undermines the arbitrariness-related motivation for adopting the Bargaining norm over calibration-based EVM norms. The arbitrariness worries concerning calibration choice are not alleviated by the Bargaining norm, rather, they are replaced by arbitrariness worries concerning the choice of a disagreement point. The Bargaining norm might initially appear to somehow allow us to eat our cake and have it too – to give us the decisiveness of a strong norm, without the arbitrariness worries that usually come with such norms. However, the arbitrariness is not eliminated, it is merely better concealed in the formalism.

Proponents of the Bargaining norm might reject this characterization by objecting to the claim that any way of fixing the disagreement point is arbitrary. They might argue that considerations of fairness, or moderation among moral theories fix the disagreement point in a non-arbitrary way. For example, perhaps fairness considerations favour the original disagreement point (a_1, a_4) in the above example, because it is composed of each theory's "worst case scenario" as a_1 is the worst alternative on v_1 and a_4 is the worst alternative on v_2.[59] On the flip side, perhaps other disagreement points can be rejected on

[59] Greaves & Cotton-Barratt (2023) call this point the "anti-utopia point" and suggest it – among several other points – as a "reasonably elegant" way of fixing the disagreement point (pp. 13-14). While they acknowledge that the lack of reference of the disagreement point is problematic, they do not fully appreciate the degree to which it undermines the original motivation for the Bargaining norm as calibration-indifferent. At some point (p. 13) they suggest fixing the disagreement point in some "reasonably simple and elegant" way and evaluating the Bargaining norm by the plausibility of its prescriptions. But this suggestion seems to reverse the direction of justification – norms typically justify their prescriptions, rather than the other way round. For example, suppose that getting ice cream has maximal expected value for me in some context, and is thus prescribed by Expected Utility theory. The fact that the theory requires that I get ice cream doesn't serve as a justification for the theory. Rather, the justification runs in the opposite direction – I am justified in getting ice cream because the theory (with its arguably plausible properties delineated in representation theorems) requires it. To draw justification for the norm from its prescriptions we would need some method, independent of the norm, for evaluating the prescriptions. In the ice cream example, we would need some way of determining that the requirement to get ice cream is plausible, independently of Expected Utility theory, and therefore the theory is justified in virtue of so requiring. In both cases – Expected Utility theory, and the Bargaining norm – it is unobvious where such a method is to be found.

fairness grounds. For example, perhaps d_1 favors v_1 and d_5 favors v_2, as they improve the "starting point" of the respective theories in the metaphoric bargaining process.

However, notice that these claims are very similar to the ones justifying the choice of a privileged calibration for strong EVM norms. There too, considerations of fairness and moderation among theories serve as a justification for favouring one way of settling IVCs over the rest. However, calibration-indifference was appealing precisely because it allowed one to arrive at decisive prescriptions without alluding to this type of justification. If an appeal to moderation and fairness among moral theories suffices to alleviate arbitrariness worries when it comes to choosing a disagreement point, then why wouldn't it suffice to alleviate such worries in relation to calibration choice? And if it does suffice in the calibration case, then the Bargaining norm's calibration-indifference ceases to be an advantage over calibration-based EVM norms. So, it seems that the Bargaining norm and strong EVM norms are more similar than we might have initially realized – both are susceptible to arbitrariness objections and might try to evade them by appealing to notions of fairness and moderation among theories. Therefore, the Bargaining norm and Strong EVM norms seem to have a similar status when it comes to arbitrariness objections.

8. Conclusion

The literature on moral uncertainty is skewed towards strong norms, treating IVC indeterminacy quite differently from other instances of indeterminacy. In this chapter I argue that this bias should be corrected, and that Weak norms ought to be taken seriously as responses to moral uncertainty with IVC indeterminacy. The problem of moral uncertainty with IVC indeterminacy is not *sui generis*, but an instance of a more general problem discussed in considerable depth in the literature, namely, choice under indeterminacy. Making progress on a particular instance of a problem is almost always best pursued by drawing on the discussion of the general one. The chapter may therefore be seen as calling for more careful and informed philosophical progress when it comes to moral uncertainty.

Beyond being good philosophical practice, taking Weak norms for moral uncertainty seriously has been shown to be fruitful in at least two ways. First, considering Weak EVM for the expectational case revealed its equivalence to non-dominance and gave rise to a new argument for the norm that does not rely on the content of the lower-order norms. Secondly, the equivalence between maximizability and non-domination sheds light on the landscape of strong norms, yielding an illuminating comparison between calibration-based and calibration-indifferent strong norms. Among other things, this comparison reveals that the Bargaining norm – a calibration indifferent norm – is more similar to calibration-based norms when it comes to arbitrariness considerations, than it initially seems.

In sum, there is much to gain from situating the problem of moral uncertainty with IVC indeterminacy in its broader context of choice under indeterminacy. While moral uncertainty might be special, it is not *that* special.

Chapter 3

Measuring the Weight of Reasons

1. Introduction

When deliberating whether to ϕ it is typical to compare reasons for ϕ-ing with reasons against ϕ-ing. We often use the term "weighing" to describe this type of comparison – we say things like "*weighing* the pros and cons", "the considerations in favour of taking the job *outweigh* those against doing so", and the like. And indeed, comparing reasons does have many of the characteristics of comparing weights – individual reasons appear to have different weights in the sense that they may count in favour of doing something to different degrees, and these weights seem to "combine" or "add up" in a way that allows comparing something like the "total weight" of the reasons for doing something, or the total degree of support they provide for doing that thing, to the "total weight" of the reasons against doing it.

This notion of the weight of reasons is also central to their relationship to other normative properties. Most centrally, facts about whether one ought to ϕ are related to facts about the relative strength or weight of the reasons for and against ϕ-ing, where very roughly, one ought to ϕ if and only if the reasons for ϕ-ing outweigh the reasons against ϕ-ing. It is controversial which of these facts – the ought fact or the weight fact – is more fundamental,[1] and which should be explained in terms of the other, but the logical relation between weights and oughts depicted in the above biconditional is close to platitudinous throughout normative theorizing. The weight-ought relationship is also supported by the phenomenology of deliberation – we typically weigh reasons for and

[1] Schroeder (2021a, 2021b) argues that reasons are more fundamental. Broome (2004) argues that ought facts are more fundamental.

against ϕ-ing with the motivation of figuring out whether we ought to ϕ. So, the goal of weighing reasons relies on some close relationship between facts about the comparative weights of reasons and facts about what one ought to do.

So, the weights of reasons are a salient part of the phenomenology of deliberation, and they play a crucial role in explaining the relations between reasons and other normative properties, most centrally – ought facts. Therefore, any theory of reasons that wishes to take these features seriously must include a notion of weight of reasons that makes sense of the ways different sets of reasons appear to have different weights in deliberation and the ways in which comparisons among such weights relate to what one ought to do (and perhaps other normative facts, e.g., facts about permissibility).

Coming up with such a notion of weight requires addressing questions of at least two types. First, analytical, or metaphysical questions: what does it mean for a reason, or a set of reasons to have certain weight? Is this notion fundamental or is it explained in terms of other things?[2] Second, structural, and measurement-related questions – what types of relations hold among different weights of reasons (or among sets of reasons with different weights)? Can weight be measured or represented numerically in a way that preserves these relations? And, if it can, what does this representation look like? The metaphysical and structural questions are somewhat, but not fully, independent. They are somewhat independent because some answers to the metaphysical questions leave the structural questions entirely open, but they are not fully independent because other answers to the metaphysical questions place constraints on, or even fully determine, the answers to the structural questions.

For example, if weight is fundamental and is not to be understood in any other terms, then the structure of this *sui generis* property is "up for grabs" – we are free,

[2] See Fogal & Risberg (2023) for an opinionated survey of views concerning these questions.

theoretically speaking, to answer the structural questions in all sorts of ways and are unconstrained by the metaphysics of weight. In contrast, if the weight of a reason is understood as the degree to which it probabilistically confirms some other proposition (Nair 2021), or the difference it makes to some conditional expected value (Sher 2019), then weight already comes with a readymade structure and numerical representation, namely that of probabilistic confirmation or conditional expected value. In these cases, answering the metaphysical questions fully fixes the answers to the structural ones. In between these two extremes, are metaphysical views that cash out weight in terms of properties which might have some structural strings attached, but whose structure and representation is not obviously fully fleshed out. For example, if weights of reasons are understood as the amounts of the normative support they provide (Fogal & Risberg 2023), then the structure of weight is the structure of normative support. While this might constrain our theorizing about the structure of weight in some ways, it does not fully determine it in the ways identifying it with probabilistic confirmation or conditional expected value does.[3]

In this chapter, I will set the metaphysical questions aside and address the structural ones. I will explore the structure of the weight of reasons and its numerical representability, by considering the types of judgments that our intuitive concept of weight licenses, while remaining largely neutral about its metaphysics. I will suggest a preliminary numerical representation, or measurement, of weight that preserves some –

[3] Another way in which the metaphysical and structural questions are not entirely independent, is that exploring the structural questions might render some metaphysical views more parsimonious, and therefore, more plausible than they would otherwise be. For example, non-reductionism about weight – the view that the weight of reasons is a primitive notion, not to be understood in terms of other notions (like probability) – would be rendered more parsimonious, and thus arguably, more plausible, if it was excused from assuming that all numerical facts about weight were fundamental. A measurement of weight – like the one proposed in this chapter – that allows deriving numerical facts from ordinal comparisons among weights would alleviate the non-reductionist from the metaphysical baggage of treating numerical facts as fundamental and would thus render the view more plausible than it otherwise would have been.

but possibly not all (see section 7) – of the relations between weights of different sets of reasons.

A numerical representation of weight could be theoretically valuable in multiple ways. First, inasmuch as it is a plausible representation, it could demonstrate the richness of the represented property. In particular, it could determine how seriously we should take the weight metaphor for reasons: physical weight has a rich numerical structure and licenses all sorts of numerical comparisons among weights; how rich is the structure underpinning the notion of weight of reasons? Does it license the same sorts of numerical comparisons? A plausible representation would allow answering these questions. Second, the representability of the weight of reasons by a rich numerical structure, may enable interesting comparisons to other numerically representable notions like probability and utility, or other types of value. The numerical structures allow potentially fruitful explorations of the relationships between weight and these possibly related notions, without reducing weight to them. So, coming up with a plausible numerical representation of weight without reducing the notion to some other numerically representable notion, allows the non-reductionist to enjoy the rich structure and its potential relation to other structures, without the metaphysical commitments of reductionism.

The remainder of the chapter proceeds as follows. In section 2, I lay out the approach I will adopt to measuring the weight of reasons – the *Representational Approach* of deriving numerical representations from intuitive qualitative properties of non-numerical structures. In section 3, I will briefly survey different types of representation with varying expressive powers that will be relevant to the weight of reasons. In section 4, I construct a structure for expressing qualitative judgments about the weight of reasons and derive a numerical representation of weight from a set of axioms constraining this structure. In section 5, I discuss the axioms consider their plausibility. In section 6, I argue for a tighter representation than the one derived. In section 7, I explore the possible shortcomings of the derived representation of weight, with respect to its ability to accommodate and justify "reason accrual" – the ways in which individual reasons contribute to the weights of collections of reasons. Finally, in section 8, I conclude.

2. The Representational Approach

My intention here is to derive a numerical representation of weight from non-numerical, and arguably more plausible, properties of weight. I begin with comparative non-numerical judgments about weights of reasons and demonstrate that if those judgments – or more precisely the facts those judgments are about – satisfy certain arguably plausible conditions, then they – or more precisely, the weight property that they refer to – may be represented numerically in a relatively unique way. The theoretical benefit of deriving a numerical "quantitative" representation from non-numerical "qualitative" features stems from the difference in plausibility between quantitative and qualitative judgments about weights of reasons.[4] To see why this might be the case, consider the following judgment forms:

(1) *The reasons in favour of ϕ-ing outweigh the reasons against ϕ-ing.*

(2) *While the reasons in favour of ϕ-ing outweigh those against it slightly, the reasons for ψ-ing outweigh the reasons against it by a lot.*

(3) *The reasons in favour of ϕ-ing are r times weightier than the reasons against ϕ-ing.* Where r is some real number.

[4] I use the term "plausibility" as an umbrella notion intended to capture a cluster of properties judgments may possess: if a judgment is plausible in the sense intended here, then it raises no, or reasonably few, qualms about its meaning, truth, metaphysics, or epistemic access. That is, its meaning is clear, the type of facts that would make it true are not mysterious, and it is not especially hard to determine whether it is true.

The first judgment form is ubiquitous and natural, and while we might have theoretical qualms about what it is for reasons to outweigh each other in this way, we have a pretty solid pre-theoretic intuitive grasp of the meaning of (1). Indeed, (1) is precisely the kind of statement we'd say as the conclusion of deliberation about whether to ϕ. The second statement might seem slightly more contrived, but is still, I think, well within the boundaries of clear pre-theoretic talk involving reasons. In contrast, the third judgment is much less natural or clear. What does it mean, one might reasonably ask, for some set of reasons to be, say, 2.4 times weightier than another set of reasons? While the meaning of (1) and (2) are clear, (3) is much more mysterious. What makes (3) true? How do we know whether (3) is true? Are there primitive facts about weight ratios for reasons that we ought to search for to verify (3)? While the properties of weight referred to in the first two judgments appear to be unsuspicious and genuine, the third judgments might appear to assume too much about the notion of weight by referring to a rigid numerical structure that isn't obviously included in it.

It is this plausibility difference between (1) and (2) on the one hand, and (3) on the other, that makes the derivation of a numerical representation of weight theoretically valuable. For, if we manage to demonstrate that qualitative judgments like (1) and (2) may be represented numerically, then we can make sense of judgments like (3) without stipulating primitive numerical notions. Quantitative judgments like (3) can be understood *in terms of* more plausible qualitative judgments like (1) and (2). And the questions of meaning, truth, metaphysics, and epistemic access regarding (3) may be answered in terms of the answers to the respective questions about (1) and (2). This way we can have a rich numerical structure for the weight of reasons, without the mysteriousness about its meaning or origin.[5]

This derivation of the representation of weight is facilitated by a theorem due to Krantz et al. (1971) from *Measurement Theory* – a field of study devoted to the derivation of numerical representations, or *measurements*, of numberless algebraic structures. The theory shows how different algebraic structures – tuples consisting of a set and some relations on it – can be numerically represented by sets of functions that map the structure's basic elements to real numbers, and the structure's relations to arithmetical relations among those real numbers.[6] The theory is useful because it allows the use of numbers to measure properties that are not initially characterized in numerical terms. In the scientific context, it provides methods to measure data from scientific observations characterized in qualitative non-numerical terms.[7] Applied to our context, the theory allows using numbers to measure the weight of reasons, a notion we might initially characterize in a more qualitative numberless way (like in judgments 1 and 2 above).

[5] This type of plausibility gap is what makes derivations of such numerical representations philosophically significant more generally. For example, in formal epistemology, representation theorems that allow deriving credence functions from certain qualitative "likelier than" relations among propositions, are philosophically useful because of the plausibility gap between the two notions. While it is not immediately obvious how a credence function represents our epistemic states (what mental state is represented by a credence function that gives *p* a credence of 0.29?), the "likelier than" relation corresponds quite naturally to a familiar epistemic state of judging one proposition likelier than another. Similarly, while it is not immediately obvious that the Kolmogorov axioms are requirements of rationality, the axioms constraining the "likelier than" relation (e.g., transitivity) are arguably intuitive as rationality requirements for our epistemic states. The theorems allow bridging both gaps – a credence function relates to our epistemic states by representing the "likelier than" relation, and the normativity of probabilism for credence emerges from the normativity of the axioms for the "likelier than" relation. See Ramsey (1931), Savage (1972), and Joyce (1999) for representation theorems in this area. See Konek (2019) for an illuminating overview of this approach.

[6] More precisely, such functions are *Homomorphisms* of the algebraic structure into a numerical structure composed of the set of real numbers and the relevant arithmetical relations on it. See Krantz et al. (1971), pp. 8-9 for a detailed discussion.

[7] For example, in the revealed preference tradition in economics, a quantitative utility function is derived from qualitative observations about preferences or choice behaviour. See Bossert & Suzumura (2010b) for a survey of this field.

3. Interlude: varieties of representation

A measurement or representation of a structure is a set of *representing functions* – functions mapping the structure's elements to real numbers. This set of functions is defined by a *representation condition* – a condition stating some relationship between the structure's relations and certain arithmetical relations among the images of the functions (the numbers the functions map the structure's elements to). The set of functions that represents the structure given a representation condition, is composed of precisely those real-valued functions that satisfy the representation condition. For example, given a structure $S = (X, \succcurlyeq)$, where X is some nonempty set and $\succcurlyeq$ is a binary relation on X, a representation of S could be defined by the following representation condition:

> **Representation condition $\boldsymbol{c}$**: the function $f: X \to \mathbb{R}$ *represents* S if and only if, for all $x, y \in X$, $x \succcurlyeq y$ if and only if $f(x) \geq f(y)$.

The representation of S, relative to this representation condition c, is then defined as the set F_c of real-valued functions, composed of precisely the functions that satisfy the representation condition c, or formally:

$$F_c \coloneqq \{f: X \to \mathbb{R} | f \text{ satisfies } c\}$$

So, F_c represents S relative to the specified representation condition c. Importantly, there are many other possible representation conditions, and thus a measurement represents a structure only relative to a representation condition. The choice of the representation condition is a matter of interest and practicality – a good representation condition should pick out a property of the structure that we wish to measure and relate it to numbers in an intuitive and useful way.

Importantly, choosing a representation condition alone doesn't ensure that it is satisfied by *any* functions. An important part of deriving a representation is to demonstrate the existence of a representing function in the sense defined by the representation condition. For instance, in the example above, if $\succcurlyeq$ is incomplete or intransitive, then the set of functions characterized by the representation condition is empty – *no* function represents S in the sense required by the representation condition. So, representation theorems always have an *existence component* – a statement of the existence of a function that satisfies the representation condition, or, in other words, that the representation F_c is not empty.

A numerical representation of a structure – the set of functions that satisfy the representation condition – gives rise to, and makes sense of, quantitative claims about the measured property (or properties). The measurement may be thought of as a structure underpinning the semantics of such claims.[8] The truth values of quantitative claims involving the measured property are determined by the set of representing functions in a supervaluationist manner – a claim is true (false) on a measurement if and only if it is true (false) on *all* of its members.[9] If a claim is true relative to some representing function but false relative to another, then it does not refer to a property represented by the measurement, and is neither true nor false relative to it. Quantitative properties that are not stable throughout the measurement are thus mere artifacts of the representation and do not correspond to any property of the represented structure.

[8] For the use of measurements as semantic structures for quantitative claims involving gradable adjectives see Kennedy (1997). Lassiter (2011) devises a measurement-based semantics for modals.

[9] I am assuming a straightforward sense of truth (and falsehood) relative to a single representing function. For example, given a measurement of weight W the claim "x is twice as heavy as y" will be true relative to $w \in W$ if $w(x) = 2w(y)$. I am assuming that the nontrivial details of the semantics that takes one from the quantitative claim to the subsequent equation, can be worked out.

On this supervaluationist picture, the expressive power of a measurement – the range of the types of claims that it makes sense of – decreases with its size.[10] As the set of representing functions grows, unanimity becomes harder to achieve and less types of judgments are rendered true or false by all of its members. On the flip side, as the representation becomes "tighter" and less functions satisfy the representation condition – so long as the measure remains nonempty – a broader range of types of statements will be decided by the functions in the set.

The size of a representation, and its corresponding expressive power, is determined by the *uniqueness component* of representation theorems – a description of the range of functions that satisfy the representation condition. The size of a representation is referred to as its *scale type* and is typically characterized by the type of transformation that holds between its elements and under which it is *closed*. A measurement F_c is *closed under a transformation* t if the t-transformations of the elements of F_c are also elements of F_c.[11] Different scale types, associated with closure under different transformations, have different expressive powers.

A measurement or representation of a structure is thus characterized by a *representation condition* and a *scale type*. The representation condition c specifies the relationships that must hold between a function and a structure for the former to represent the latter, and defines the set of representing functions F_c. The scale type specifies the size and expressive power of the representation in terms of the type of transformation relation that holds among its members and under which it is closed. More generally, and more formally, a representation is characterized by a representation

[10] Different representations are typically infinite sets of equal cardinality, so my use of the term 'size' here is loose rather than the standard set-theoretic notion. For example, if F_1 is a proper subset of F_2, I take F_1 to be a "smaller" or "tighter" representation than F_2 in this loose sense, even if the two representations have equal cardinality.

[11] This property is often referred to as "uniqueness up to a transformation".

condition c and a scale type s associated with a transformation t_s, in the following manner:

> Let $S = (X, R_1, .. R_n)$ be a structure where X is a set and $R_1, \dots, R_n$ are relations on X. Given a representation condition c, the set of functions F_c *c-represents* S on a s-scale if and only if it satisfies the following requirements:
>
> 1. **Existence**: F_c is not empty.
> 2. **Transformation**: any two elements of F_c are t_s-transformations of each other.
> 3. **Closure**: F_c is closed under t_s-transformation.

Three scale types, each associated with a transformation type, will be of use for the purposes of this chapter. They are defined as follows:

1. **Ordinal Scale**: a representation represents a structure on an *ordinal scale* if it satisfies Transformation and Closure with respect to *monotone transformation*:
   A function f' is a *monotone transformation* of a function f iff there is a monotone (order preserving) function t such that $f'(x) = t\big(f(x)\big)$ for all $x \in X$. Or for short $f' = t(f)$.

2. **Interval Scale**: a representation represents a structure on an *interval scale* if it satisfies Transformation and Closure with respect to *positive affine transformation*:
   A function f' is a positive *affine transformation* of a function f iff there exist real numbers $a > 0$ and b such that $f'(x) = af(x) + b$. Or for short $f' = af + b$.

3. **Ratio Scale**: a representation represents a structure on a *ratio scale* if it satisfies Transformation and Closure with respect to *proportionality transformation*:
   A function f' is a *proportional transformation* of a function f iff there exists a positive real number $a > 0$ such that $f'(x) = af(x)$. Or for short $f' = af$.

The transformations appear in an order of strength – if t is a proportional transformation, then it is a positive affine transformation, and if it is a positive affine transformation then it is a monotone transformation. Therefore, the scale types appear in an ascending order of expressive power – interval scales are strictly more expressive than ordinal scales, and ratio scales are strictly more expressive than interval scales. To see this, consider the following types of statements with respect to weight, and the subsequent table:

1. **Ordinal**: the weight of x is greater than the weight of y.
2. **Interval**: the weight difference between x and y is r times the weight difference between z and w, where r is a real number.
3. **Cardinal**: x weighs much more than y (but z weighs only slightly more than w).
4. **Ratio**: the weight of x is r times the weight of y, where r is a real number.
5. **Additivity**: the weight z equals the sum of the weights of x and y.

Recall that a representation makes sense of a statement only if all of its members agree on its truth value. In the above three scales, this unanimity condition is equivalent to truth-value robustness with respect to the relevant transformation. If a claim's truth value relative to a representation function is robust with respect to the scale's transformation, then it will have the same truth value throughout the representation. The following table lists the types of statements whose truth value is determined by each of the three scale types:

TABLE 3.1: *Type of statements determined by the three scale types*

Statement type/Scale type	**Ordinal**	**Interval**	**Ratio**
Ordinal	V	V	V
Interval		V	V
Cardinal[12]		V	V
Ratio			V
Additivity			V

So, the expressibility relations between the three scales are strict. A ratio scale is strictly more expressive than an interval scale as it makes sense of all the statement types accommodated by an interval scale, as well as other statement types that are undetermined by interval scales. Similarly, an interval scale is strictly more expressive than an ordinal scale. These differences in expressibility will be consequential in the discussion below of the type of scale most appropriate for measuring weight of reasons.

4. Deriving a representation of weight

I now turn to the central task of the chapter – deriving a numerical representation of the weight of reasons. As mentioned above, my strategy is to construct an algebraic structure that allows expressing qualitative judgments about the weights of reasons, and then to employ the resources of the theory of measurement to derive a numerical representation of weight from constraints, or axioms, on this structure. I divide this task into two stages.

[12] I take cardinal statements to be about relative weight differences – whether a weight difference is considered "slight" or "vast" is determined relatively, in relation to other weight differences. Therefore, I take this type of statement to be settled by ratios of weight differences, and thus, by both interval and ratio scales. To see that such claims are supported by interval scales, consider the following claim about temperature (or heat), a property measured on an interval scale: "Cairo is much hotter than Edenborough but only slightly hotter than Athens".

First, I present a model that enables reasoning about reasons and the weight relations that might hold between them in a precise way. Second, I construct the qualitative algebraic structure from which a numerical representation of weight is ultimately derived. I do this by gradually going through different types of ordinal comparative judgments about weight and expressing them within the structure. I begin with basic weight comparisons (like (1) above), and subsequently move on to comparisons of "outweighings" – pairs of sets of reasons where one element of the pair outweighs the other, (like (2) above). The emergent structure enables the application of a representation theorem and the derivation of a numerical representation of the weight of reasons.

4.1. The Objects of Weight

I have thus far been talking quite loosely about reasons as the objects of weight but haven't yet characterized them precisely. Reasons, in the normative sense I am interested in here, are considerations in favour of actions or attitudes. That it will rain is a reason for taking an umbrella; that Deborah's phone is on her desk is a reason to believe that she hasn't gone home yet; that the lion escaped the zoo is a reason to fear leaving the house. Reasons are therefore always related to some action or attitude of which they count in favour – they are always reasons *for* taking a certain action or adopting a certain attitude. The scope of action and attitude types that reasons may apply to is vast – there could be reasons to assert, believe, fear, hope, like, hate, enjoy, worry, take delight in, be humoured by, and any other action or attitude for which it makes sense to talk about reasons for its adoption.

Reasons can also be considerations against some action or attitude ϕ. However, I will think of such reasons as reasons *for* an *omissive* action or attitude – for not taking the action or adopting the attitude that the reason is a consideration against, or for short:

reasons for $\neg\phi$. So, reasons will be treated as unipolar in my framework – as only counting in favor of actions/attitudes. The notion of counting against will be captured by omissive or negative actions and attitudes of the form $\neg\phi$ which represent "not ϕ-ing".

Reasons are also related to agents. That the local pub is screening the Tottenham game may be a reason for David, but not for Joel to go there, if for example, David is an avid Spurs fan and Joel doesn't care for football. Perhaps some reasons are *agent-neutral* – perhaps some reasons count in favour of some actions or attitudes for *all* agents, but this is not always the case.[13] So, in the general case, reasons should be thought of in relation to both actions or attitudes and agents. A reason is therefore always a reason *for* an agent *to* carry out an action or adopt an attitude.

Finally, reasons are related to possible worlds. Perhaps in the actual world, that it will rain is a reason for you to take an umbrella, but in some close-by possible world in which it does not rain, or in which it is so hot you'd prefer to get wet, you no longer have this reason to take an umbrella. As this example demonstrates, something can be a reason for an agent to act in a certain way in one possible world, but not in another.[14] So, reasons should be thought of in relation to actions or attitudes, agents, and possible worlds.

[13] See Parfit (1984) for an early discussion.

[14] The world-relativity of reasons may also play a role in explicating Schroeder's (2008, 2009) distinction between *objective* and *subjective reasons*. Objective reasonhood is a factive relation – objective reasons to ϕ are true propositions that count in favour of ϕ-ing. In contrast, subjective reasons to ϕ are propositions the agent believes but need not be true. For example, if Bernie believes that his glass contains gin and tonic, he may have a subjective reason to take a sip (if, e.g., he likes gin and tonic) even if he is wrong and in fact the glass contains gasoline (Schroeder 2008, p. 60). In such a case the proposition "the glass contains gin and tonic" would not be an objective reason for Bernie to take a sip – as it is false – but it *would* be a subjective reason for him to do so. While a precise characterization of subjective reasons in elusive, it will likely involve whether a proposition is an objective reason in some other possible world. The normative import of the proposition "the glass contains gin and tonic" (and that Bernie believes it), is plausibly related to it being an objective reason for Bernie to take a sip in some non-actual possible world. If we take the weight of subjective reasons at a world to be related to the weight of the corresponding objective reasons (possibly at a different world), then the representation of weight of objective reasons derived below could also be used to construct a representation of the weight of subjective reasons.

Therefore, a reason is always a reason *for* an agent *to* take an action or adopt an attitude *at* a possible world.

I will think about reasons as true propositions and represent them as such in the model.[15] However, the model may be interpreted more ecumenically with respect to views regarding the ontology of reasons. In particular, the model, and the derivation of a measurement of weight if facilitates, may be useful to any view that allows *associating* reasons with true propositions, even if it does not *identify* reasons as such entities. For example, if reasons are facts,[16] properties,[17] or mental states,[18] then they may straightforwardly be represented by corresponding true propositions describing the relevant considerations, facts, properties, or mental states, and much of the model can be applied with such correspondences in mind. In any event, I will refer to reasons as true propositions, and leave any alternative interpretations of the model implicit.

On this picture, a reason is a proposition that stands in a certain relation – let us call it the *reason relation* – to an agent, an action or attitude, and a world.[19] We can thus define the *reason relation* as a four-place relation that holds between a proposition p, an agent i, an action or attitude ϕ and a possible world w, when p is a reason for i to ϕ at w. To model this notion formally, let W, A and I be nonempty sets of possible worlds, actions or attitudes, and agents respectively. For simplicity, I will assume the sets are finite. Propositions, as is standardly the case, will be represented by the sets of possible worlds (subsets of W) at which they are true. Let the reason relation R be a four-place relation

[15] I will thus be thinking about *objective* reasons in Schroeder's (2008c) sense. However, if, as suggested in the previous footnote, subjective reasons can be understood in terms of objective reasons in other possible worlds, then the model below could also be used to express relations among subjective reasons, and for measuring their weight.

[16] Dancy (2000), Scanlon (2014).

[17] Nagel (1970).

[18] Davidson (1963). For other views on the ontology of reasons see Schroeder (2021b), Chapter 2.

[19] Scanlon's (2014) "reason relation" replaces the world variable with a "circumstances" variable. Here I abstract away all intra-world circumstance variation.

among propositions, agents, actions/attitudes, and worlds such that $R(p, i, \phi, w)$ encodes that p is a reason for i to ϕ at w.

Different types of reasons give rise to different reason relations. For example, a proposition may be a prudential reason, but not a moral reason, for i to ϕ at w. To accommodate this fact there must be at least two distinct reason relations – a prudential one and a moral one. Similarly, offering someone some money for believing a proposition may give rise to a prudential reason, but not an epistemic reason for them to do so, thus giving rise to a third reason relation for epistemic reasons. A full taxonomy of the different types of reason relations would likely be complex and controversial, and it is beyond my scope here. I will simply note that my discussion here should be understood as conducted relative to a single reason relation. The formal apparatus suggested here and the derivation of a numerical representation of weight to follow, may be applied to any type of reasons separately – prudential, moral, epistemic, etc. – but not to more than one type at once. There is no obvious, or apparently useful universal weight scale that applies to all types of reasons. Rather, the weight of different types of reasons should be conceived of and measured separately.

A core part of practical reasoning, and more generally, of deliberation over whether to ϕ, consists in figuring out which propositions are reasons to ϕ (and which are reasons not to ϕ) – discerning the considerations for (and against) ϕ-ing. In terms of the reason relation, searching for the reasons for ϕ-ing, is searching for the propositions that bear the reason relation toward the agent, ϕ, and the world that the agent inhabits. To express this formally in our model, let r be a *reason function* – a function that takes as its input a set of propositions X, an agent i, an action/attitude ϕ, and a world w, and yields a subset of X consisting of precisely those elements of X that are reasons for i to ϕ at w. r is defined in terms of R: for any set of propositions X, agent i, action/attitude ϕ, and world w: $r(X, i, \phi, w) = \{p \in X | R(p, i, \phi, w)\}$. Recall that reasons are assumed to be true propositions, and so the set of reasons for an agent to ϕ at a world, will always be a subset of the set of propositions that are true at that world. For a possible world w, let $[w]$ denote the set of all the propositions that are true at w, i.e., all supersets of $\{w\}$.[20] So, $r([w], i, \phi, w)$ is the set of *all* reasons for i to ϕ at w.

On this picture, the weight of a reason should be thought of as a property of not merely a proposition, but a property of an element of the reason relation – a proposition, an agent, an action/attitude, and a world: a tuple of the form $t = (p, i, \phi, w)$ such that $t \in R$. To generalize this to sets of reasons, the weight of a set of propositions should be thought of as a property of a tuple of the form (X, i, ϕ, w) where X is a set of propositions that are reasons for i to ϕ at w.[21] So, the objects of weight, and the relata of the "weightier than" relation among sets of reasons (about which I elaborate below), are tuples of this kind.

I will make two simplifying assumptions about these tuples for the remainder of the chapter. First, I will abstract away the agent variable, and assume that they are constant throughout the discussion.[22] And so, the relata of the "weightier than" relation in the discussion below are triples of the form (X, ϕ, w). I do this to avoid the entanglements of interpersonal comparisons of weight of reasons, e.g., judgments of the form: "the reasons in favor of Jack ϕ-ing are weightier than the reasons in favor of Jill ψ-ing". While we might ultimately want to make sense of such comparisons, at least in some cases, their meaning is not as straightforward as comparisons involving a single agent. I also suspect that making sense of interpersonal weight comparisons would require careful consideration of some of the thorny issues raised by interpersonal welfare or

[20] Formally: $[w] \coloneqq \{p \subseteq W | w \in p\}$.

[21] Formally, $r(X, i, \phi, w) = X$. Or equivalently: for all $p \in X$, $(p, i, \phi, w) \in R$.

[22] I will often omit the background agent altogether and say things like "a reason to ϕ at w", which should be understood as "a reason *for the agent* to ϕ at w".

utility comparisons, and this would require careful and separate treatment that is beyond my scope here.[23] So, the derived numerical representation below is *intrapersonal* – it measures the weight of reasons for an individual agent across actions and worlds. However, if one is willing to incorporate interpersonal weight judgments of the above form, the derivation may be easily amended to give rise to an interpersonal representation of weight.

Second, I will assume that all weight-related judgments are unaffected by propositions that are not reasons. So, for example, the "weightier than" relation among sets of reasons to be considered below treats the triple (X, ϕ, w) and the triple $(r(X, \phi, w), \phi, w)$ identically. This relation, and any other weight-related judgment to be considered, simply ignores the members of X that are not reasons. This simplifying assumption obviates the need to assume, as was stated above, that all propositions in X are reasons to ϕ at w. On this assumption, the objects of weight are simply triples of the form (X, ϕ, w), where X is some set of propositions, which may or may not be reasons to ϕ at w.

We are now in a position to construct the structure expressing comparative judgments about the weight of reasons, and to derive a measurement of weight from it.

4.2. Derivation

The most basic type of judgment that a notion of weight should accommodate, is a "weightier than" comparison among the reasons for and against an action or attitude. That is, judgments of the form: "the reasons for ϕ-ing outweigh the reasons against ϕ-ing". As pointed out above, this type of judgment is part and parcel of practical

[23] This is especially likely for theories that link reasons to desires, e.g., Schroeder (2007). For some challenges and possibilities in relation to interpersonal comparisons of utility see List (2003) and Bradley (2008).

deliberation – agents trying to figure out whether they ought to ϕ typically arrive at conclusions of this form (or of the inverse form: "the reasons for ϕ-ing are outweighed by the reasons against doing so"). Accommodating this type of judgment is also crucial for reasons to be related to what one ought to do in the way outlined above: if the reasons for ϕ-ing outweigh the reasons against ϕ-ing then you ought to ϕ. For reasons to be related to "ought facts" in this way, and for one notion to be explained in terms of the other, our notion of weight must make sense of this type of comparative judgment. In the terminology of our model this, "weightier than" comparison is a comparison between two triplets: $([w], \phi, w)$ and $([w], \neg\phi, w)$ – between the reasons to ϕ at w and the reasons to not ϕ at w.

But "weightier than" judgments don't only hold among *universal sets of reasons* – sets that include *all* the reasons to ϕ at w – they also hold among smaller sets. You've been offered a job – the fact that the colleagues are nice is a reason to accept the offer, and the fact that the commute is long is a reason to reject the offer. The first reason is weightier than the second (and this may be so even if there are many other reasons that bear on the decision). These types of comparisons may give rise to (or may be analysed in terms of) "qualified" ought judgements – if we restrict attention to these two considerations, then you ought to take the job.[24] While the example involves comparison of two reasons, similar comparisons involving sets of reasons (subsets of the universal set of reasons) sound similarly natural. In terms of our model, this is a comparison between (X, ϕ, w) and $(Y, \neg\phi, w)$, where X and Y are arbitrary sets of propositions.

"Weightier than" judgments may also hold between reasons for the same action or attitude. For example, the fact that the colleagues are nice and the fact that there's free

[24] This qualified or restricted understanding of "ought" is very much in the spirit Kratzer's (1981) analysis of "ought" and other modals. In her analysis, the truth value of "ought" statements is determined relative to a "modal base" – a contextually determined body of information.

coffee are both reasons for accepting the job offer, but the first reason is weightier than the second. The example includes a comparison of two individual reasons, but similar comparisons among sets of reasons are similarly intuitive. For example, the facts that the commute is pleasant and that there's free coffee, taken together, might be weightier – as reasons for accepting the job offer – than the fact that the colleagues are nice. In our model, this is a comparison between (X, ϕ, w) and (Y, ϕ, w).

More generally, "weightier than" judgments may also hold between reasons for compatible actions. For example, the fact that it's sunny out is a reason to apply sunscreen, and the fact that it's hot out is a reason to get ice cream, where it is possible – let us assume – to do both. Still, it makes sense to say that the fact that it is sunny is a weightier reason to apply sunscreen than the fact that it is hot is to buy ice cream (sunscreen will prevent skin diseases, ice cream is just a tasty cold treat). Here too, the example involves a comparison between two singletons, but the general type of judgment involving sets of reasons is similarly plausible. In our model, this is a comparison between (X, ϕ, w) and (Y, ψ, w), where ϕ and ψ are compatible.

Finally, "weightier than" judgments hold across worlds. For example, the fact that it's sunny outside is a reason to apply sunscreen. As it happens, it is not hot out, but had it been hot out, that (counterfactual) fact would have been a reason to get ice cream. The weight of these two reasons is comparable, even though they inhabit different possible worlds. It is reasonable, for example, to claim that the fact that it's sunny is a weightier reason to apply sunscreen in the actual world than the counterfactual fact that it is hot is a reason to get ice cream in the relevant counterfactual world. Or, that the fact that it is sunny is a weightier reason to apply sunscreen than the counterfactual fact that it is hot – if it were hot – would be a reason to get ice cream. In our model, this is a comparison between (X, ϕ, w) and (Y, ψ, w').

If all the above judgment types make sense, then our notion of weight allows an extensive range of comparisons under the "weightier than" relation. The relation appears to hold between a broad range of different sets of reasons, actions or attitudes, and worlds. To represent this relation within our model, let T be the set of triples of the from (X, ϕ, w) where X is a set of reasons, ϕ is an action/attitude and w is a world.[25] And let $\succcurlyeq$ be a binary relation over T such that $(X, \phi, w) \succcurlyeq (Y, \psi, w')$ means that the reasons in X to ϕ at w are at least as weighty as the reasons in Y to ψ at w'. We can then define the "strictly weightier than" relation $\succ$ and the "equally weighty" relation $\sim$ in the usual ways.[26]

The set of triples T taken together with the "weightier than" relation $\succcurlyeq$ form an algebraic structure, which may in principle be represented numerically. One would hope to find a plausible set of conditions to place on this structure that would be sufficient for the derivation of a weight function that represents $\succcurlyeq$ in an appropriate sense. However, while some possibly intuitive conditions for $\succcurlyeq$ may come to mind, the structure $(T, \succcurlyeq)$ alone is not rich enough – as far as I can see – to give rise to an adequate measure of weight. If $\succcurlyeq$ is assumed to be complete and transitive, then straightforwardly, an ordinal scale representation of weight may be derived. But such a representation would lack the expressive power required of an adequate measurement of weight – it would not make sense of cardinal weight claims and would not accommodate claims involving degrees of weight. So, to derive an adequate representation of weight, we must enrich the underlying structure further by considering an additional type of judgment about weight, namely, comparisons of the severity of outweighing among sets of reasons.[27]

[25] Formally, $T = \mathcal{P}(\mathcal{P}(W)) \times A \times W$, where $\mathcal{P}$ is the powerset operator: $\mathcal{P}(W)$ it the set of all propositions and $\mathcal{P}(\mathcal{P}(W))$ is the set of all sets of propositions.

[26] For any two triples $t_1, t_2 \in T$, (a) $t_1 \succ t_2$ iff $t_1 \succcurlyeq t_2$ and $\neg(t_2 \succcurlyeq t_1)$, and (b) $t_1 \sim t_2$ iff $t_1 \succcurlyeq t_2$ and $t_2 \succcurlyeq t_1$.

[27] Two other options may come to mind for enriching the structure to enable a derivation of a more informative representation. First, following Von Neumann and Morgenstern (1947) we might introduce the notion of *mixtures* of sets of reasons – probability distributions over T – and derive an interval-scale

Our intuitive notion of weight licenses not only ordinal comparisons of weights of reasons, but also comparisons of the strength or severity of outweighing relations among reasons. For example, that I will receive a "kidney donor" T-shirt (p), that my friends will admire me (q), and that I will save someone's life (r), are all reasons to donate a kidney. Furthermore, r outweighs q and q outweighs p. So far this is all expressible in terms of the "weightier than" relation. But, while q outweighs p *somewhat*, r outweighs q by *a lot*. In even weaker terms, r outweighs q *by more*, or *more severely* than q outweighs p. This judgment is an ordinal comparison not among weights, but among outweighing relations – the first outweighing relation is stronger, or more severe, than the second. In this example, the severity of outweighing comparison involves reasons for a single action at a single world. However, it is plausible that wherever it makes sense to talk about outweighing, it also makes sense to talk about severity of outweighing. Therefore, the extensive applicability of the "weightier than" relation demonstrated above implies a similarly extensive applicability of the severity of outweighing relation.

This type of judgment – comparisons of severity of outweighing – may be modelled by a binary relation among *pairs* of triplets: let $\succcurlyeq^*$ be a binary relation among pairs of elements of T,[28] such that for any four triples $a, b, c, d \in T$, $ab \succcurlyeq^* cd$ holds iff the reasons in a outweigh those in b at least as severely as the reasons in c outweigh those in

representation of the "weightier than" relation over mixtures of T given certain axioms. However, this is not a promising direction. The notion of a mixture of reasons and the idea of ordering such mixtures with respect to weight is odd and unnatural – it is not clear what such mixtures would amount to and what would it mean for such a mixture to be weightier than another. Second, we might add a *concatenation operation* on T to the structure and use the representation theorems for *Extensive Measurement* in Krantz et al. (1971), to derive a ratio scale measurement of weight. A *concatenation operation* is a function ($\circ: T \times T \rightarrow T$) that maps pairs of triples to their "sum" a third triple that represents the reasons in both triples taken together. The challenge with this representational strategy – and the reason why I opt for a third strategy in the text – is that concatenation might not make sense for many pairs of triples (imagine concatenating across actions and worlds). See section 7 for an additional discussion of this option.

[28] Equivalently, $\succcurlyeq^*$ can be thought of as a 4-place relation over T such that $\succcurlyeq^* \subseteq T \times T \times T \times T$.

d. We can define the symmetric $\sim^*$ and asymmetric $\succ^*$ parts of $\succcurlyeq^*$ standardly.[29] Also, the "weightier than" relation $\succcurlyeq$ can be defined in terms of $\succcurlyeq^*$ as follows: for any two triples $a, b \in T$, $a \succcurlyeq b$ if and only if $ab \succcurlyeq^* aa$. That is, a outweighs b if and only if, the reasons in a outweigh the reasons in b at least as severely as the reasons in a outweigh the reasons in a. Since the reasons in a are equally weighty as themselves, this condition amounts to the reasons in a being at least as weighty as the reasons in b, which is precisely the $\succcurlyeq$ relation we are out to define.

In contrast to the "weightier than" relation $\succcurlyeq$, the severity of outweighing relation $\succcurlyeq^*$ – or more precisely, the structure $(T, \succcurlyeq^*)$ – *does* allow a derivation of a potentially adequate representation of weight. To see this, consider the following conditions, or "axioms", on the structure $(T, \succcurlyeq^*)$ and the subsequent theorem:

1. **Weak order**: $\succcurlyeq^*$ is a complete and transitive over pairs of elements of $T \times T$.
2. **Invertibility**: if $ab \succcurlyeq^* cd$ then $dc \succcurlyeq^* ba$.
3. **Monotonicity (of severity)**: if $ab \succcurlyeq^* a'b'$ and $bc \succcurlyeq^* b'c'$ then $ac \succcurlyeq^* a'c'$.
4. **Richness**: if $ab \succcurlyeq^* cd \succcurlyeq^* aa$ then there exists $e, e' \in T$ such that $ae \sim^* cd \sim^* e'b$.

Theorem 3.1 (Krantz et al. 1971): If $\succcurlyeq^*$ and T satisfy the above four axioms then there exists a weight function $w: T \to \mathbb{R}$ such that for $a, b, c, d \in T$, $ab \succcurlyeq^* cd$ if and only if $w(a) - w(b) \geq w(c) - w(d)$. Furthermore, w is unique up to positive affine transformation, so if w' has the above property then there exist real numbers $a > 0$ and b such that $w' = aw + b$.[30] [31]

[29] For any four triples $a, b, c, d \in T$, (a) $ab \succ^* cd$ iff $ac \succcurlyeq^* cd$ and $\neg(cd \succcurlyeq^* ab)$, and (b) $ab \sim^* cd$ iff $ab \succcurlyeq^* cd$ and $cd \succcurlyeq^* ab$.

[30] See Krantz et al. (1971), p. 151. The original theorem includes a fifth "Archimedean" axiom that is automatically satisfied when T is finite. So, adding this axiom – which regulates certain infinite sequences of triples – would allow one to derive a numerical representation of weight for the non-finite case as well.

[31] In the terms introduced in section 3, the derived representation is characterized by the following representation condition c: the function $w: T \to \mathbb{R}$ *represents* $S = (T, \succcurlyeq^*)$ if and only if, for all $a, b, c, d \in T$,

So, if the severity of outweighing relation $\succcurlyeq^*$ satisfies the above four axioms, then the weight of reasons is numerically representable on an interval scale. Notice that while the relata of $\succcurlyeq^*$ are *pairs* of triples, the objects that the weight function assigns numbers to are simply triples – so the additional structure required for expressing comparisons among outweighing relations, allowed the derivation of an interval scale numerical representation of weight. Of course, since the relata $\succcurlyeq^*$ are interpreted as weight differences between triples, w allows expressing both weight and weight differences.

Before examining the properties of this representation, let us consider the underlying axioms.

5. The Axioms

5.1. Weak Order

Weak order requires that the severity of outweighing relation be complete and transitive. Completeness means that all weight differences are comparable, or that any two outweighing relations are comparable with respect to severity. Transitivity means that if a first outweighing relation (between two sets of reasons) is more severe than a second, which itself is more severe than a third, then the first relation is also more severe than the third. In simpler terms, Weak order requires that it be possible to place outweighing relations neatly in an order of severity (with possible ties).

On the face of it, Weak order is a natural property for a relation like severity of outweighing. In other contexts, severity of outweighing, defeating, or of other

$ab \succcurlyeq^* cd$ if and only if $w(a) - w(b) \geq w(c) - w(d)$. Theorem 3.1 states that F_c satisfies *Existence* and *Positive affine Transformation*. F_c satisfies *Positive affine Closure* because the representation condition is robust to positive affine transformation, and so if w_1 satisfies c and w_2 is a positive affine transformation of w_1, then w_2 also satisfies c and thus $w_2 \in F_c$.

comparative relations, appear to satisfy Weak order – for example, we can order pairs of suitcases with respect to severity of physical outweighing, or can order football matches with respect to severity of defeat using goal differences. And in many cases, it might very well be possible to order outweighing relations among sets of reasons with respect to severity. Indeed, the above kidney-donation example appears to be such a case, and it is easy to see how it may be extended to include other pairs of sets of reasons.

However, on a more careful consideration, there is good reason to think that Weak order does not hold in full generality in the case of outweighing relations among sets of reasons. In contrast to comparisons of physical weight or goal differences, that are straightforwardly measurable and one-dimensional, comparing sets of reasons often requires comparing different types of values and considerations. For example, when deliberating between two job offers, the reasons in favour of each will often involve a plurality of considerations and values – location, pay, prestige, interest, colleagues, promotion prospects, coffee, dress code, etc. While comparing the jobs with respect to any one of these considerations might be straightforward – e.g., the pay-related reasons to accept one of the offers might be obviously weightier than the pay-related reasons to accept the other – it might not be obvious how to compare the two jobs with respect to the full scope of relevant considerations. In other words, it might be unobvious which set of reasons is weightier – the reasons for accepting one job offer or the reasons for accepting the other.

One way of taking the non-obviousness of the above type of comparison seriously, is to accept that in some cases, there is no determinant fact of the matter as to which of two sets of reasons is weightier – for some pairs of sets of reasons it is neither the case that one outweighs the other, nor is it the case that they are equally weighty. In other words, the "weightier than" relation among sets of reasons is incomplete. And if the "weightier than" relation is incomplete then so is the severity of outweighing relation –

if there is no outweighing relation between some pair of sets of reasons, then that pair of sets cannot bear any severity of outweighing relation to another pair of sets of reasons.[32]

This view, that the "weightier than" relation among sets of reasons is incomplete, is closely related to views regarding the incompleteness of betterness and preference relations. It has been quite broadly argued that for some pairs of actions or alternatives, it is neither the case that one is better than the other nor is it the case that they are equally good. Similarly, for some pairs of actions or alternatives, and some agents, it is neither the case that the agent prefers one alternative to the other, nor that they are indifferent between them. The incompleteness of betterness or preference may entail – given some plausible assumptions about how these notions relate to reasons – the incompleteness of the "weightier than" relation among sets of reasons. If for example, ϕ-ing is better than ψ-ing if and only if the reasons in favor of ϕ-ing outweigh those in favor of ψ-ing, then incompleteness in betterness entails incompleteness in the "weightier than" relation. Similarly, if ϕ-ing is preferable to ψ-ing if and only if the reasons in favor of ϕ-ing outweigh those in favor of ψ-ing, then incompleteness in preference entails incompleteness in the "weightier than" relation.[33] Notice that these biconditionals are consistent with a variety of metaphysical views regarding the relation between reasons, betterness and preference, including views that take betterness and preference to be explained in terms of reasons,[34] views that take reasons to be explainable in terms of

[32] More precisely, the incompleteness of $\succcurlyeq^*$ is entailed by the incompleteness of $\succcurlyeq$ and Invertibility. If the "weightier than" relation $\succcurlyeq$ is incomplete, then there are triples $a, b \in T$ such that neither $a \succcurlyeq b$ nor $b \succcurlyeq a$. Using the definition of $\succcurlyeq$ in terms of $\succcurlyeq^*$, this entails that neither $ab \succcurlyeq^* aa$ nor $ba \succcurlyeq^* aa$. It then follows from Invertibility that neither $ab \succcurlyeq^* aa$ nor $aa \succcurlyeq^* ab$, and thus that $\succcurlyeq^*$ is incomplete.

[33] The reasons that relate to preference in this way may be of a different type than the reasons that relate to betterness in the corresponding biconditional. For example, perhaps the preference related reasons are prudential whereas the betterness related reasons are moral. So, the incompleteness of preference and betterness may entail incompleteness in different types of "weightier than" relations that hold among different types of reasons.

[34] On a *reasons-fundamentalist* view, according to which all normative properties are explainable in terms of reasons, betterness is explainable in terms of reasons, and inasmuch as preference is a normative notion, so

betterness or preference,[35] and views that take all three notions to be explained in terms of some other notion. Regardless of the underlying priority relations, if one takes reasons to closely correlate with betterness or preference, then incompleteness in the latter can entail incompleteness in the "weightier than" relation.

However, even if we accept that the "weightier than" relation is incomplete, and therefore that the severity of outweighing relation is too incomplete, we might still wish to accept Weak order as a local constraint or as a tentative idealization. Accepting Weak order as a local constraint that applies to some proper subset of triples would allow deriving a measure of weight for the sets of reasons that are fully comparable. This could be thought of as a numerical representation of the weight of reasons for the "ordinary" cases, those that do not involve any tricky incommensurability. While being less theoretically ambitious, such a measure would still apply to a broad range of weight judgments. Perhaps more usefully, accepting Weak order as a tentative idealization would allow ignoring the complexities of incomparability to derive a preliminary measure of weight that misrepresents the weights of incomparable sets of reasons, but yields important insights about weight relations, nonetheless. This preliminary idealized representation can also be a useful starting point for devising more realistic and less idealized representations that incorporate incompleteness.[36]

is preference. For a defence of reason-fundamentalism see Schroeder (2021a, 2021b). For an interpretation of preference as stemming from reasons see Dietrich and List (2013).

[35] Lord & Maguire (2016) mention such a view (p. 18). Finlay (2014, 2019) understands reasons in terms of explanation of goodness, and Maguire (2016) understands reasons in terms of promotion of value. Both views can be naturally extended to understanding outweighing relations in terms of betterness or preference relations (or related concepts). For criticism see Fogal & Risberg (2023).

[36] One promising way forward would be to apply generalizations of decision theory that relax the completeness axiom for preference (e.g., Aumann 1962, and Galaabaatar and Karni 2013) to the current context. Such an application would allow, under certain conditions, deriving a family of representations (each corresponding to a different way of completing the incompleteness in $\succcurlyeq^*$). Claims about weight would then be adjudicated relative to this family of representations in a supervaluationist manner,

5.2. Invertibility

Invertibility regulates the extension of the severity relation to "negative" outweighing relations. Until now, my discussion of comparisons of outweighing relations with respect to severity, implicitly presupposed that all outweighing relations were "positive", in the sense that when it is said that a outweighs b more severely than c outweighs d, it is implicitly assumed that a outweighs b and c outweighs d. Indeed, severity comparisons are most natural in such cases. However, this assumption is inessential, and comparisons of severity may be extended to apply to cases in which the outweighing relations are "negative", that is, to cases involving pairs of alternatives ab where b outweighs a.[37]

Invertibility requires that the order of severity of negative outweighing relations be the inverse of the order of severity of the corresponding positive relations. So, for example, if ab and cd are both positive (a outweighs b and c outweighs d), and $ab \succcurlyeq^* cd$, then dc and ba which are both negative, are ordered inversely, i.e., they satisfy $dc \succcurlyeq^* ba$. This is very natural if we think of pairs of triples representing differences of weight – if the difference between the weight of a and the weight of b is positive, then the difference between the weight of b and the weight of a is negative, and the two differences will have inverted relationships to other differences. However, if we think of $\succcurlyeq^*$ more qualitatively as an order of severity of outweighing, talk of negative outweighing is less natural. Nonetheless, the introduction of the notion, and its regulation by Invertibility appear innocuous enough. One may think of the negative side of $\succcurlyeq^*$ as a mere formal construction, defined as the mirror image of its positive side.[38] Since the negative side of

determining all claims involving weight relations that are unaffected by incompleteness, while leaving claims involving incommensurable triples indeterminate.

[37] Formally, the outweighing relation between $(a, b) \in T \times T$ is *positive* iff $ab \succcurlyeq^* aa$, and *negative* iff $ba \succcurlyeq^* aa$. Invertibility ensures that when (a, b) is positive, (b, a) is negative: $ab \succcurlyeq^* aa \succcurlyeq^* ba$.

[38] For an illuminating discussion of the use of such formal constructs, that do not represent real objects, in the context of decision theory, see the correspondence between Aumann and Savage in Dreze (1990). Savage compares these non-representing formal entities to construction lines in geometry (p. 78).

the relation is entirely definable in terms of its positive side, there is no room for arbitrary decisions about "negative differences" to influence the derived representation of weight.

5.3. Monotonicity (of Severity)[39]

Monotonicity is a condition that applies to what I will term *connected pairs of pairs*. A pair of pairs is *connected* if the second element of the first pair is identical to the first element of the second pair, i.e., pairs of the from: ab, bc where $a, b, c \in T$. Monotonicity requires that we treat – in a certain sense – the pair ac as the sum of the connected pairs ab and bc with respect to severity of outweighing. So, loosely speaking, if a outweighs b to some degree, and b outweighs c to another degree, then a outweighs c to the sum of these degrees. Of course, this talk of degrees and sums of degrees is figurative, as the algebraic structure includes neither notion. Instead, this loose idea of summing, or adding up, is expressed by the relation between the connected pair ab, bc and another (strategically chosen) connected pair $a'b', b'c'$. In particular, $a', b', c' \in T$ are chosen such that $ab \succcurlyeq^* a'b'$ and $bc \succcurlyeq^* b'c'$. So, instead of saying that a outweighs b to a certain degree of severity, we say that a outweighs b at least as severely as a' outweighs b', and similarly, b outweighs c at least as severely as b' outweighs c'. In other words, $a'b'$ and $b'c'$ are stand-ins for the notion of the degree to which a outweighs b and the degree to which b outweighs c, respectively. Monotonicity then requires that whenever this relation holds between the two connected pairs, then the respective "sums" ac and $a'c'$ satisfy $ac \succcurlyeq^* a'c'$. That is, the "sum" of the severity of a outweighing b and the severity of b outweighing c is at least as great as the respective sum involving a', b' and c'.

[39] I call this axiom "Monotonicity of severity" to distinguish it from "Monotonicity of weight" discussed in section 7. My use of the ambiguous term "Monotonicity" in this subsection should be understood as shorthand for "Monotonicity of Severity".

Is Monotonicity plausible? Is it plausible to think of the severity of the outweighing in ac as something like the sum of the severities in ab and bc? To try to answer these questions, let us consider some arguably intuitive properties of connected pairs through the kidney donation example mentioned above. The example included the following three reasons to donate a kidney:

(1) I will save a life (p).

(2) My friends will admire me (q).

(3) I will get a "kidney donor" T-shirt (r).

Let us assume – quite plausibly, I think – that p outweighs q and q outweighs r, and notice that pq and qr are a connected pair.[40] What, if anything, does this entail regarding the relation between p and r as far as outweighing goes? Firstly, I think it's fairly clear that if p outweighs q and q outweighs r then p must outweigh r. It would be absurd to say, given the way we set up the example, that the fact that I will save a life is not a weightier reason than the fact that I will receive a T-shirt. More generally, this type of example suggests that the "weightier than" relation among sets of reasons is transitive. While the transitivity of the "weightier than" relation is entailed by Monotonicity – indeed p outweighing r is necessary for thinking of pr as the sum of pq and qr with respect to severity of outweighing – it is strictly weaker than it.[41]

But more can be said about the outweighing relations among these three reasons. If p outweighs q and q outweighs r, then q can be thought of as having a weight that lies

[40] I will use the proposition letters ambiguously, to refer both to the propositions themselves, and to their respective triples. The different uses should be distinguishable by the context. So, for example, when I say that p outweighs q, I mean that (p, ϕ, w) outweighs (q, ϕ, w) where ϕ is the action of donating a kidney, and w is some fixed background world. Also, to minimize notation, I will drop the singleton brackets and use (p, ϕ, w) instead of $(\{p\}, \phi, w)$ when the set of propositions includes only one element.

[41] The transitivity of the "weightier than" relation may be stated in terms of $\succcurlyeq^*$ and connected pairs as follows: for any triples $a, b, c \in T$, if $ab \succcurlyeq^* aa$ and $bc \succcurlyeq^* aa$ then $ac \succcurlyeq^* aa$. Stated in these terms, transitivity of the "weightier than" relation straightforwardly follows from Monotonicity.

somewhere *between* the weights of p and q. This betweenness can be understood in different ways. A weak sense in which q may be said to lie between p and r – let us call this sense *Weak Betweenness* – is that p outweighs r more severely than the outweighings in both pq and qr. To see why Weak Betweenness is plausible, let us compare the degree or severity to which saving a life outweighs gaining admiration, to the degree or severity to which saving a life outweighs receiving a T-shirt – all as reasons to donate a kidney. It would be odd if the former outweighing were more severe than the latter, for to move from saving a life to receiving a T-shirt one must travel a greater distance, in terms of weight, than the move from saving a life to gaining admiration. Gaining admiration is "on the way" – with respect to weight – from saving a life to receiving a T-shirt. And so, it seems reasonable to require, as Weak Betweenness does, that $pr \succcurlyeq^* pq$. A similar process would demonstrate the reasonableness of requiring that p outweigh r more severely than q outweighs r, and thus $pr \succcurlyeq^* qr$. Like transitivity of the "weightier than" relation, Weak Betweenness is entailed by, but still strictly weaker than, Monotonicity.[42]

Monotonicity alludes to a stronger sense of betweenness according to which the outweighing in pr is not only more severe than the outweighings in pq and qr, but it is tantamount to their sum, in the manner explicated above. To make this sense of betweenness more concrete, let us add a corresponding connected pair to our example. Let us assume that the following three propositions are reasons for me to go to the park:

(1) I'll get to see my friends (u).
(2) I'll get to enjoy the sunshine (v).
(3) I'll get to eat ice cream (w).

[42] Proof: from Weak Order $pq \succcurlyeq^* pq$ and by assumption $qr \succcurlyeq^* qq$, so by Monotonicity $pr \succcurlyeq^* pq$. Similarly, by assumption $pq \succcurlyeq^* qq$ and by Weak Order $qr \succcurlyeq^* qr$, so by Monotonicity $pr \succcurlyeq^* qr$.

Let us assume that while u outweighs v and v outweighs w, these outweighings are moderate – all three propositions are of similar strength as reasons to go to the park. Let us now compare the connected pair uv, vw to the connected pair pq, qr, and let us assume that $pq \succ^* uv$ and $qr \succ^* vw$. This is plausible if we accept that in contrast to the marginal weight differences among the reasons to go to the park, p vastly outweighs q (saving a life is way stronger of a reason than gaining admiration), and q considerably outweighs r (the admiration of one's friends is considerably more valuable than a T-shirt). Monotonicity then requires that $pr \succ^* uw$ – that p outweigh r more severely than u outweighs w.[43] That is, Monotonicity requires pr to relate to uw as the sum of pq and qr would relate to the sum of uv and vw, if each of these pairs was associated with a number in a way that satisfies the antecedent of Monotonicity (with $\succcurlyeq^*$ replaced by $\geq$). This is a stronger condition than Weak Betweenness – not only does q lie *between* p and r, but the distance (with respect to weight) between p and r can be thought of as *divided* by q into two intervals pq and qr the sum of which is tantamount to pr (and similarly for u, v and w).

Monotonicity seems plausible in the above example, and I take the notion of sums of outweighings among connected pairs to be fairly natural. However, one might be sceptical of the amount of structure Monotonicity imposes on the severity of outweighing relation, especially when it comes to thinking of summing the severity of outweighing relations. The weaker conditions entailed by Monotonicity on the other hand – transitivity, and Weak Betweenness – may invoke less scepticism, as they do not help themselves to a summation notion. Still, even if one is ultimately reluctant to accept

[43] This strict version of Monotonicity – if $ab \succ^* a'b'$ and $bc \succ^* b'c'$ then $ac \succ^* a'c'$ – follows from Monotonicity and Invertibility. Proof: assume (1) $ab \succ^* a'b'$ and (2) $bc \succ^* b'c'$ and assume by way of negation (3) $a'c' \succcurlyeq^* ac$. (1) and Invertibility entail (4) $b'a' \succcurlyeq^* ba$. (4) and (3) entail by Monotonicity $b'c' \succcurlyeq^* bc$ which is in contradiction to (2). Therefore, $ac \succ^* a'c'$.

Monotonicity, it may nonetheless be considered a useful starting point for exploring the derivation of weight measures that rely on weaker conditions.

5.4. Richness

Richness requires that the set of triples T include enough elements to allow *dividing* severe outweighing relations into less severe ones. Let us say that a triple e *divides* a positive pair of tripes ab, if e lies *between* a and b with respect to weight in the sense alluded to in the discussion of Monotonicity above: $a \succcurlyeq e \succcurlyeq b$.[44] That is, the weight difference between a and b can be thought of as divided by e into two intervals – ae and eb. Let us call such a triple a *divider.* Richness requires that for any two positive pairs ab and cd such that $ab \succcurlyeq^* cd$, we can find in T (a) a divider e that is outweighed by a as severely as d is outweighed by c: $ae \sim^* cd$, and (b) a divider e' that outweighs b as severely as c outweighs d: $e'b \sim^* cd$.

It is not obvious that every pair of triples can be divided in such a way. Perhaps reasons come in a limited number of weights and aren't sufficiently fine-grained. However, reasons related to gradable values or variable quantities, like monetary reasons, could potentially give rise to at least some of the dividers required by Richness. Reasons of the form "I will receive x dollars if I ϕ" would give rise to quite a fine-grained array of weights. Such reasons could be used to divide certain positive pairs. For example, both of the following propositions are reasons to go bowling:

(1) I will get a workout (p)

(2) I will get an adrenaline rush (r)

And let us assume that p outweighs r. The following reason is a good candidate for a divider of pr:

[44] Or equivalently, in terms of $\succcurlyeq^*$: $ae \succcurlyeq^* aa$ and $eb \succcurlyeq^* aa$.

(3) I will get an adrenaline rush and one dollar (q)

Plausibly, q lies between p and r with respect to weight, and different versions of q involving different numbers of dollars would divide pr in different ways. However, this suggestion suffers from two problems. First, it puts more pressure on the plausibility of Weak Order, and specifically on the plausibility of completeness. For, the demandingness of richness increases with the size of T, and extending T to include a plethora of somewhat contrived reasons like (3), might render completeness harder to accept. Indeed, the paradigmatic cases where the completeness of betterness or preference appears to fail are insensitive to *mild sweetening*. That is, when two alternatives are unrelated by preference, this does not change when one of them is mildly improved or *sweetened*, by, e.g., adding a small amount of money to it.[45] So, extending T to include reasons like (3) may result in generating more pairs of triples that are unrelated by the "weightier than" relation.

The second problem with this strategy of adding gradable reasons to T, is that it might be limited to the case of reasons for action. Small monetary values can amount to reasons to act in certain ways, but they cannot typically amount to reasons for belief and other attitudes, at least not the right kind of reasons. For this strategy to work beyond reasons for action, we must find analogous sweeteners for belief and other attitudes, and it may not be obvious how to do so.[46]

However, even if this suggestion of gradable reasons is ultimately rejected, T can be extended artificially to satisfy Richness. We may treat the dividers as mere formal constructs that facilitate the derivation of a weight measure, and then ignore their weights in the derived measure. The derived weight measure would still adequately represent the properties of the structure which correspond to genuine normative properties, like the

[45] See Raz (1986) Ch. 13, Chang (1997), and Hare (2010).

[46] For sweetening in the epistemic case – "evidential sweetening" – see Schoenfield (2012).

relations between non-artificial triples. Echoing the discussion of Invertibility, here too, we may treat some parts of the formal model as not corresponding to anything of normative import, and still reap the theoretical benefits of its output (with some adjustments).

In sum, the four underlying axioms on the severity of outweighing relation that allow the above derivation are plausible to different degrees. While Invertibility and Richness may be seen as mostly formal conditions that introduce innocuous formal constructs, Weak Order and Monotonicity are substantive conditions and should invoke more scepticism especially in relation to issues of comparability and "summation" of severity. However, even if the more substantive axioms are ultimately rejected, I think they are useful starting points for alternative derivations. The axioms and the theorem may therefore be treated as a reference point from which one may deviate to come up with other, perhaps looser, representations of weight.

6. Tightening the Representation

Let us now return to the representation side of Theorem 3.1 – the derived weight function. The derived representation measures weight on an interval scale, i.e., the weight function is unique only up to positive affine transformation. Therefore, the derived measure allows making sense of ordinal, cardinal, and interval claims. That is, it makes sense of claims like "the reasons for ϕ-ing outweigh the reasons for ψ-ing", "the reasons for ϕ-ing are *much* weightier than the reasons against ϕ-ing" and "the weight difference between the reasons for ϕ-ing and the reasons against ϕ-ing is twice as great as the weight difference between the reasons for ψ-ing and the reasons against ψ-ing". The measure therefore makes sense of much of our theorizing concerning weights of reasons: it doesn't merely license ordinal outweighing claims, rather, it also allows making more nuanced distinctions between reasons that outweigh other reasons to different degrees – some outweigh slightly, others vastly outweigh.

However, there are reasons to aspire to a tighter representation of the weight of reasons, in particular, one that measures weight on a ratio scale. The main difference between a ratio scale and an interval scale is the designation of an essential zero point. Interval scales are unopinionated regarding units *or* zero points – temperature, for example, can be measured by scales that vary in both units and zero points. In contrast, ratio scales include fixed zero points – for example, while different scales for measuring physical weight have different units, having a weight of zero is a unit-independent property – weighing zero kilograms is the same as weighing zero pounds. In contrast, having a temperature of zero is not unit-independent – having a temperature of 0°C and having a temperature of 0°F are distinct properties.

The zero point in a ratio scale serves as an origin, or reference point relative to which the relevant objects are measured with respect to the represented property. This structure allows for a broader range of quantitative judgments than those expressible by an interval scale. First, a ratio scale makes sense of *ratio claims* – statements about the ratios of measured quantities (as opposed to statements about ratios of *differences* in such quantities). For example, it makes sense to say that Jack is twice as tall as his son, because height is measured on a ratio scale with a fixed zero point (of 0 cm). In contrast, it doesn't make obvious sense to say that Jack started eating twice as late as his son did. This is because clock time is plausibly measured on an interval scale and choice of zero point is arbitrary. To make sense of lateness ratios like the one above we'd need to fix a third time relative to which the events of Jack starting eating and his son starting eating may be measured.[47] The second type of judgment expressible by ratio scales but not by interval

[47] I take this comparison between height and clock time from Lassiter (2011), Chapter 2.

scales, are *additivity claims* – claims that refer to adding up different quantities of the measured property by summation. For example, while it makes sense to say that the height of Jack is the same as the sum of the heights of his two children, it does not make sense to say that the temperature in Jerusalem on some summer day equals the sum of the temperatures in London and Copenhagen.[48]

The weight of reasons is a good candidate to be measured on a ratio scale for two reasons. First, it has a natural zero point in the empty set of reasons, and second, at least on some views, it licenses additivity claims.

It is quite plausible, given our intuitive notion of weight, to think of empty sets of reasons as having zero weight. More precisely, it is natural to think of *null triples* – triples of the form $(\emptyset, \phi, w)$[49] – as having zero weight. One necessary condition for identifying null triples as occupying the zero point for weight is that they are comparable to other triples with respect to weight. Of course, in *some* sense, empty sets of reasons have *no* weight, as there are no reasons in them to bear any weight. But there must be something to say about the weight of empty sets of reasons, as they appear to be comparable to non-empty sets with respect to weight. Consider for example the claim, that if there are *some* reasons to ϕ and no reasons not to, then one ought to ϕ. This claim has the same structure and function as ordinary outweighing judgments, and it relates a reason comparison to an "ought conclusion" in much the same way as such judgments. What makes it the case

[48] It is of course possible that on a certain summer day the temperature in Jerusalem (35°C) is the sum of the temperatures in London (20°C) and Copenhagen (15°C). But this relationship disappears when we switch to a Fahrenheit scale on which the temperatures in Jerusalem, London, and Copenhagen on that day would be 95°F, 68°F and 59°F respectively. Sums of temperatures – and any other property measured on an interval scale – are thus inessential artifacts of the scale one chooses to represent them in.

[49] The formalism as constructed includes both triples of the form in the text as well as triples of the form $(\{\emptyset\}, \phi, w)$. That is, it includes both triples with an empty set of propositions, and triples with a set of propositions that includes only the empty set. Both types of triples should be thought of as null. Since the weight relations ($\succcurlyeq$ and $\succcurlyeq^*$) are assumed to ignore propositions that are not reasons (see section 4.1), they should treat $(\{\emptyset\}, \phi, w)$ and $(\emptyset, \phi, w)$ equivalently, as the empty set is true at no worlds and is therefore never a reason.

that you ought to ϕ – it seems reasonable to say – is that the reasons for ϕ-ing outweigh the reasons against it (even though, there are no such reasons). So, much like physical weight, when it comes to reasons, no weight is comparable to some weight.

Another necessary condition for identifying null triples as a zero point for weight is that they all have the same weight. If there are no reasons to ϕ and there are no reasons to ψ, then the weight of the reasons in favor of each action is the same. Again, on some level it is funny to talk about the weight of "the reasons" when there are none, but ϕ-ing and ψ-ing have the same normative status, and they both bear similar relations to alternative actions that have some reasons in their favor. These normative properties are, I think, most naturally explained in terms of weight of reasons – the reasons for both actions have no weight.[50]

Finally, null triples are a natural zero point because they are a natural minimum point for weight, and thus a natural benchmark for measuring the weight of nonempty sets of reasons. A set of reasons cannot have less weight than no weight – it can count in favour of ϕ-ing to different degrees, the minimum of which, is not counting in favor of ϕ-ing at all.[51] Therefore, the weight or reasons for ϕ-ing in some nonempty set may be naturally thought of in terms of the weight added by those reasons to the minimum weight – the weight provided by no reasons. In other words, like physical weight, weight

[50] Also, if, as suggested above, this lack of weight is comparable to other none-null weights, then there should be no problem comparing the reasons for ϕ-ing and the reasons for ψ-ing with respect to weight.

[51] You could of course think of weights has having a sign – when the reasons for ϕ-ing have positive weight they count in favour of ϕ-ing, and when they have negative weight, they count against ϕ-ing. However, on my unipolar setting, reasons against ϕ-ing are simply reasons in favor of another action, namely, not ϕ-ing. So, the reasons for ϕ-ing have minimal weight when they don't count in favor of ϕ-ing, regardless of whether they count against ϕ-ing. Whether they count against ϕ-ing can be determined by considering the same propositions in relation to not ϕ-ing.

of reasons appears to have an essential minimum, and the weight of a set of reasons is naturally thought of in terms of the weight-difference between it and the minimum.[52]

In sum, null triples satisfy two necessary conditions for having zero weight – they are comparable to other non-null triples, and they all have the same weight – and they have the characteristic "reference point" property of zero points. That is, it is natural to think of the weights of nonempty sets of reasons as being measured against the empty set. So, the weight of reasons has a natural zero point in the empty sets of reasons.

The second reason weight is a good candidate for ratio-scale measurement is that a ratio scale would allow for neutrality with respect to the additivity of the weight of reasons. When deliberating whether to ϕ we typically find ourselves "tallying" or "adding up" reasons for ϕ-ing and reasons against ϕ-ing and comparing the relative total weights or strengths of the reasons on each side. A common metaphor for this process is that of placing objects on the two sides of a scale and comparing how they balance. The reasons for ϕ-ing are placed on one side and the reasons against ϕ-ing are placed on the other, and the weightier set of reasons ultimately tilts the scale for or against ϕ-ing. One way of taking this metaphor and the phenomenology of adding up reasons seriously, is to endorse *additivity* for the weights of reasons. Additivity is the view that, much like physical weight, the weight of a set of reasons is equal to the sum of the weights of its elements.

Additivity is a strong and controversial view, and there are weaker views that take reasons to be only *partially additive* – some, but not all, sets of reasons are such that their weight is the sum of the weights of their elements – as well as views that reject additivity

[52] Note that designating the empty set of reasons as the zero point does not prevent other, nonempty sets of reasons from having zero weight as well. I am not alone in setting the zero point of weight in this way. C. Brown (2014), Sher (2019), and Nair (2021) all set the zero point as the weight of "non-reasons" (Brown 2014, p. 789) – propositions that do not count in favour (or against) the relevant action.

altogether. Representing weight on a ratio scale leaves all these views on the table – it leaves room for filling in the details in ways that render reasons additive, partially additive or non-additive. In contrast, representing weight on an interval scale is inconsistent with additivity or partial additivity, as relations among sums are not represented by interval scales (because they are not robust to positive affine transformation). So, the only way to represent weight without deciding whether, and to what degree it is additive, is to do so on a ratio scale.

So, if we are to express the idea that empty sets of reasons have zero weight, and leave open the possibility of additivity, we ought to represent weight on a ratio scale. One way of deriving a ratio scale is by fixing a zero point on an interval scale – any designation of a certain triple as having zero weight would narrow down the set of representing functions from an interval scale to a ratio scale. In some cases, such a designation may be objectionably arbitrary, as it is carried out by numerical fiat and does not track any property of the underlying algebraic structure. However, designating the empty set as the zero point for the weight of reasons is not arbitrary, as it expresses properties of the notion of weight – namely, the above properties that render the empty set a natural zero point – even if those properties are not represented in the algebraic structure.

Tightening the representation by designating all null triples as having a weight of zero requires imposing a condition that ensures that all such triples have equal weight. To do so, let $E \subset T$ be the set of all null triples, and let us add the following condition to the four axioms above:

Null-equivalence: if $e, e' \in E$ then $ee' \sim^* ee$.

Null-equivalence requires that the weight difference between any two null triples be equal to the weight difference between a triple and itself, which, in quantitative terms, is zero. To express the idea articulated above, that null triples are minimal with respect to weight, let us state the following further condition:

> **Minimality**: if $e \in E$ then for all $a \in T$, $ae \succcurlyeq^* aa$.

Minimality requires that the weight difference between any triple and a null triple is non-negative – that a triple outweighs a null triple at least as severely as it outweighs itself. Since Minimality entails Null-equivalence,[53] we can simply add Minimality to the four original axioms to arrive at the following corollary:

> **Corollary 3.1**: if $\succcurlyeq^*$ satisfies the four axioms above and Minimality, then there exists a weight function $w: T \to \mathbb{R}$ such that
>
> (a) for $a, b, c, d \in T$, $ab \succcurlyeq^* cd$ if and only if $w(a) - w(b) \geq w(c) - w(d)$,
> (b) if $e \in E$ then $w(e) = 0$,
> (c) and for all $a \in T$, $w(a) \geq 0$.
>
> Furthermore, w is unique up to proportional transformation, so if w' has the above properties then there exists a real number $a > 0$ such that $w' = aw$.[54]

The tighter representation derived in corollary 3.1 represents weight on a ratio scale, and thus makes sense of a broader range of judgments about weight. In particular, it allows making sense of ratio claims and additivity claims, and thus leaving open the question of weight additivity. However, the resulting ratio scale has a different representational status than the original interval scale (derived in Theorem 3.1), as it is

[53] To prove this, first let us prove a claim I will term *Twins*: for all triples $a, b \in T$, $aa \sim^* bb$. Proof: from completeness $aa \succcurlyeq^* bb$ or $bb \succcurlyeq^* aa$. If $aa \succcurlyeq^* bb$ then by Invertibility $bb \succcurlyeq^* aa$, and if $bb \succcurlyeq^* aa$ then by invertibility $aa \succcurlyeq^* bb$. So, either way, $aa \sim^* bb$. Now, to prove that Minimality entails Null-equivalence, assume Minimality and let $e, e' \in E$ be arbitrary null triples. Minimality entails that $e'e \succcurlyeq^* aa$. From Twins and transitivity, it follows that $e'e \succcurlyeq^* ee$, and by invertibility it follows that $ee \succcurlyeq^* ee'$. Minimality also entails that $ee' \succcurlyeq^* aa$ and analogous reasoning leads to $ee' \succcurlyeq^* ee$. So $ee \sim^* ee'$ as required by Null-equivalence.

[54] See appendix B for poof.

not derived from purely qualitative properties of the underlying structure. Moving from the original interval scale to the ratio scale is facilitated by the numerical fiat that the weight of null triples equals zero – a constraint that does not correspond to any qualitative property of the underlying algebraic structure. However, if as argued above this choice of the zero point represents a genuine property of the weight of reasons – even if we failed to represent this property in the algebraic structure – then perhaps worries of arbitrariness can be fended off. Still, it would be preferable if the zero point for weight was derived from a corresponding qualitative property in the algebraic structure, instead of being imposed numerically on the representation. I take this to be worthy of further exploration in future work.

7. Things left to be desired: Reason Accrual

The derived representation gives rise to a notion of weight that accommodates much of our theorizing about reasons. It makes sense of the idea that sets of reasons can outweigh each other, and that outweighing comes in degrees – reasons can slightly outweigh, and they can vastly outweigh other reasons; it accommodates judgments about the relative weight of reasons – some reasons have little weight, others are extremely weighty, and it allows talking about the sum of the weights of different sets of reasons (in some cases). Finally, the derived representation allows understanding strong numerical claims involving ratios and sums of weights in terms of more intuitive qualitative judgments about the severity of outweighing among sets of reasons.

However, the representation, and the underlying algebraic structure, do not include any constraints on the relations between the weights of sets of reasons and the weights of their subsets. The subset relation is not referred to by the axioms, and therefore, sets of reasons are treated by the representation as entirely independent of their subsets when it comes to weight. As a result, the model expresses no necessary relation between the weight of a set of reasons and the weights of any of its subsets, including the singleton subsets composed of individual reasons. By remaining silent on the set-subset relationship, the model therefore may be understood as an expression of the following view about the relationship between weights of different sets of reasons:

> **Independence**: The weight of a set of reasons is independent of the weights of its (proper) subsets.

Independence takes the weights of sets of reasons to be determined independently of each other, and thus the weight of each set of reasons is determined in isolation – it is unconstrained by the weights of its elements or any of its subsets.[55] Independence may be motivated by *Holism* (Dancy 2004) – the view that reasons are context-sensitive in the sense that the weight of a reason – and indeed, whether a proposition is a reason at all – depends on the context in which it is considered. If sets of reasons are understood as part of the context, then a set of reasons and any of its subsets inhabit different contexts and are thus entirely independent of each other – indeed the same proposition may cease to be a reason when considered as a member of a different set. If Holism is true, and sets of reasons are contextual parameters, then the weight of a reason when considered with other reasons, is distinct from, and independent of, the weight of a reason when considered in isolation.[56] Therefore, on such a picture, Independence is true.

[55] Weights are of course constrained by the axioms, but these do not allude to the subset relation and thus do not constrain relations between sets and their subsets *as such*. The axioms could *happen* to constrain the relationship between sets and their subsets, but only due to other relations between them. For example, transitivity could require that for some sets of reasons X and Y where $Y \subseteq X$, X is weightier than Y. But this would not be because of the subset relation but because of "weightier than" relations that hold between X, Y and other sets of reasons.

[56] Given a triple (X, ϕ, w) where X is not a singleton, Holism entails that the weight of some proposition $p \in X$, is distinct from, and independent of, the weight of the triple (p, ϕ, w). The formal framework I adopt here doesn't have the expressive means to talk about "the weight of p" in the context of the triple (X, ϕ, w)

The view that is the negation of Independence – let us term it *Dependence* – includes many sub-views that vary in the type of constraints they place on the relation between weights of sets of reasons and weights of their subsets. For example, views that reject Independence may take the weights of individual reasons to merely constrain, to partially determine, or to fully determine the weights of sets composed of them. In what follows I will explore some central ways in which independence may be false, and their implications for a representation of the weight of reasons.

Independence may be challenged by allusion to cases in which individual reasons appear to "add up" when they are considered together. In such cases there's an intuitive sense in which individual reasons *contribute to* the weight of a set of which they are members, such that the weight of a set of reasons is at least as great as the weight of each individual reason in the set. To see the intuitiveness of this picture, consider the following example from Nair (2016):

> *"There is a movie theater and a restaurant across town. And suppose that in order to get to that side of town I must cross a bridge that has a $25 toll. The toll is a reason not to cross the bridge. The movie is a reason to cross the bridge and the restaurant is also a reason to cross the bridge. It may be that if there were just the movie to see, it wouldn't be worth it to pay the toll and if there were just the restaurant, it wouldn't be worth it to pay the toll. But given that there is both the movie and the restaurant, it is worth it to pay the toll."*[57]

In Nair's example, the weight of the set of reasons composed of both the reason provided by the movie (m) and the reason provided by the restaurant (r) is strictly greater than the weight of each of those reasons considered alone. The weight of each reason appears to somehow contribute to the weight of the set composed of both. So, it would

without resorting to the triple (p, ϕ, w) and thus shifting the context. See C. Brown (2014) for a framework that allows reasoning about weights of reasons in a context-dependent setting.

[57] Nair (2016), p. 56. Also appears in Nair (2021). The examples in this section are all taken from Nair (2021) and some of the discussion of the scope of Additivity in this section follows his.

be wrong to say that the weight of the set $\{m, r\}$ is independent of the weights of m and r. Assuming that the relationship between sets and individual reasons exhibited in the above example holds more generally, something like the following condition may be endorsed:

> **Weak Monotonicity (of weight)**:[58] if $a, b \in T$ such that $a = (X, \phi, w)$, $b = (p, \phi, w)$ and $p \in X$ then $w(a) \geq w(b)$.[59]

Weak Monotonicity requires that the weight of a set of reasons be at least as great as the weight of any of its elements as seems to be suggested by the above example. A slightly stronger condition in the same spirit would require that the weight of a set of reasons be at least as great as the weight of any of its subsets:

> **Strong Monotonicity (of weight)**: if $a, b \in T$ such that $a = (X, \phi, w)$, $b = (Y, \phi, w)$ and $Y \subseteq X$ then $w(a) \geq w(b)$.

Strong Monotonicity entails Weak Monotonicity, as the singletons of the elements of X are subsets of X. Both conditions are versions of Dependence as they require that weights of sets of reasons be constrained by the weights of their elements or subsets, and thus amount to a rejection of Independence. As mentioned above, neither the derived representation nor its underlying axioms discussed in the previous sections incorporate any condition constraining the relation between weights of sets and subsets of reasons, and thus the derived measurement of weight satisfies neither Weak nor Strong Monotonicity. However, the axioms are consistent with these conditions, and thus if we express the Monotonicity conditions in qualitative terms, it might be possible to tighten

[58] I use the term "Monotonicity of weight" to distinguish it from the axiom discussed above (Monotonicity of severity). Use of the term "Monotonicity" in this section is to be understood as shorthand for "Monotonicity of weight".

[59] By taking Weak Monotonicity to be a generalization of the dynamic in Nair's example, I am assuming that the weight p contributes to $w(a)$ is aptly captured by $w(b)$. That is, I am rejecting Holism, or the view that triples are contextual parameters (or both).

our representation further in a way that satisfies them. The following conditions are the qualitative analogues of Weak and Strong Monotonicity:

> **Weak Monotonicity (qualitative)**: if $a, b \in T$ such that $a = (X, \phi, w)$, $b = (p, \phi, w)$ and $p \in X$ then $ab \succcurlyeq^* aa$.
>
> **Strong Monotonicity (qualitative)**: if $a, b \in T$ such that $a = (X, \phi, w)$, $b = (Y, \phi, w)$ and $Y \subseteq X$ then $ab \succcurlyeq^* aa$.

If we add either of these conditions to the five axioms used to derive the representation, we might be able to derive a weight function that satisfies the Monotonicity conditions, and thus entails that in the above example the weight of $\{m, r\}$ is greater than the weights of either of its elements. So, at least *prima fascia*, it appears that our model can be revised to accommodate this version of Dependence.

But Nair's example can be thought to motivate a stronger version of Dependence. While both Weak and Strong Monotonicity entail that the weight of $\{m, r\}$ is at least as great as the weights of its elements, neither condition provides a detailed account of *how* the weights of the individual reasons contribute or give rise to the weight of the set of which they are members. If one thinks a stronger, determining, relationship holds between the weights of individual reasons and sets of reasons, one might be tempted by a condition termed *Additivity*, on which the relationship between sets of reasons and individual reasons is simply additive – the weight of a set of reasons is the sum of the weights of its elements:

> **Additivity**: if $a \in T$ is a triple of the form (X, ϕ, w) then $w(a) = \sum_{p \in X} w(p, \phi, w)$.

Additivity is a strong condition, as it renders the weights of all sets of reasons as fully determined by the weights of their elements. It also entails Strong Monotonicity and thus Weak Monotonicity, as a sum of positive numbers is always at least as great as any of them.[60] Additivity is tempting because it accommodates the interaction between reasons that appears to be exhibited in Nair's example in a simple way. It provides a straightforward account of the sense in which individual reasons *contribute* to the weight of a set of which they are members and an explanation for *why* sets of reasons are weightier than their elements – namely, the weights of the members are simply summed up to give rise to the weight of the set. Additivity may also be appealing if one takes the weighing metaphor sufficiently seriously. If the weighing of reasons for and against an action has the same structure as physically weighing objects on a scale, then Additivity is true, as physical weight is additive.[61]

Additivity is a special case of a broader view we may refer to as *Atomism*, that takes the weights of sets of reasons to be fully determined by the weights of their elements.[62] Additivity requires this determination to take the form of summation, but it could, in principle, take other forms. For example, a view that takes the weight of a set of reasons to be equal to the product of the weights of its elements would also be a version of Atomism. More generally, Atomism is the view that the weights of all sets of reasons are expressible in terms of the weights of individual, or *atomic*, reasons.

[60] Additivity is stated as a constraint on the relation between the weight of a set and the weights of its elements, but it entails the following constraint I term *Partition*, between the weight of a set and the weights of its subsets. **Partition**: if $a \in T$ is a triple of the form (X, ϕ, w), and π is some partition of X then $w(a) = \sum_{Y \in \pi} w(Y, \phi, w)$. Additivity entails Partition because Additivity entails that the weight of each element of a partition is the sum of the weights of *its* elements. Therefore, the sum of the weights of the elements of a partition of X equals the sum of the weights of the elements of X. Given Partition it is easy to see that Additivity entails Strong Monotonicity, and not merely Weak Monotonicity.

[61] For accounts sympathetic to Additivity see Berker (2007), Lord & Maguire (2016), and Wedgwood (2022). For a recent argument against it see Keeling (2023). Sher (2019) and Nair (2021) discussed below argue that weight of reasons is only *partially additive* – weights satisfy Additivity only under certain conditions.

[62] See C. Brown (2014) who states this as a supervenience condition: the weight of nonatomic reasons supervene on weights of atomic reasons. Brown's framework is different from mine – among other things, it allows for context-sensitivity – but the notion of Atomism is expressible in both frameworks.

There are two potential ways to accommodate Additivity in a representation of the weight of reasons. First, we might derive a limited representation of weight only for individual reasons, and then extend it to sets of reasons by additivity – setting the weights of sets of reasons as the sums of the weights of their elements. So, instead of applying the axioms above to the set of *all* triples T, one may apply them to a restricted set of *atomic triples* $T_A \subseteq T$ consisting of precisely those triples whose first element – the set of propositions – is a singleton.[63] If the restriction of $\succcurlyeq^*$ to $T_A \times T_A$ satisfies the above axioms, then an *atomic weight function* w_A from T_A to non-negative real numbers (and unique up to proportional transformation) may be derived. The weights of sets of reasons may then be equated to the sums of the weights of their elements as given by w_A. A full weight function w mapping *all* triples to non-negative real numbers may then be defined in terms of the atomic weight function and additivity as follows:

Extended Atomic Weight: For all triples $a = (X, \phi, w) \in T$, $w(a) = \sum_{p \in X} w_A(p, \phi, w)$.

Notably, this strategy would work to accommodate other forms of Atomism that express the weight of a set as a different function of the weights of its elements.[64]

This strategy suffices to accommodate Additivity (or other forms of Atomism) but not to justify it. While the atomic weight function is derived from purportedly plausible constraints on the severity of outweighing relation among atomic reasons, the extension of this functions to other triples, and Additivity, are not. Rather, Additivity is imposed as a numerical fiat on the relationship between sets of reasons and individual reasons and

[63] Formally, $T_a = \{(X, \phi, w) \in T : |X| = 1\}$.

[64] We can think of different versions of Atomism as each providing some *extension function* for the weight of reasons – a function that takes an atomic weight function and returns an extension of it to all triples. The strategy in the text to accommodate Additivity is composed of two stages – first, the derivation of an atomic weight function and second, the extension of that function to all triples. The second stage will therefore change when the strategy is applied to accommodate other versions of Atomism, as such versions will differ in their extension functions.

does not arise from any qualitative constraints on the underlying structure. So, if one was hoping to use the derivation to justify an additive representation of weight, then this strategy of extension by additivity will not do.

The second way to accommodate Additivity in a representation of weight is to enrich the underlying algebraic structure in a way that allows expressing the notion of "adding up" reasons or considering two (or more) reasons together. One initially promising way to go about this is to introduce a *concatenation* operation $\circ$ to the algebraic structure, so that for two triples $a \circ b$ expresses the reasons in both triples "taken together". For example, in Nair's example, if $a = (m, \phi, w)$ represents the reason the movie provides for crossing the bridge and $b = (r, \phi, w)$ represents the reason provided by the restaurant for doing so, then $a \circ b = (\{m, r\}, \phi, w)$ represents the two reasons taken together. Krantz et al. (1971) have several representations theorems for such structures that give rise to additive measurements.[65] So, inasmuch as the axioms of those theorems are convincing when applied to weight of reasons, they may be instrumental in deriving an additive representation of weight. The central complication that would have to be addressed under this approach is the scope of concatenation – it may not be plausible to "add up" the reasons (with respect to weight) in *any* pair of triples (think, e.g., of cross-action and cross-world concatenation). Therefore, the precise scope of concatenation and its implications for representability would have to be explored by such an attempt.[66]

However, even if tighter representations that incorporate Additivity or Monotonicity are achievable, neither condition appears to hold universally. In particular,

[65] See Krantz et al. (1971), chapter 3: *Extensive Measurement.*

[66] Concatenation is a function that maps pairs of triples to their sums, which are also triples: $\circ: T \times T \to T$. Adding such a function to the structure therefore imposes a richness requirement on T – it must include triples that represent the sums of all pairs of its elements.

in some cases sets of reasons appear to be *less* weighty than any of their elements. To see this, consider the following example from Prakken (2005):

> *"Consider by way of example two reasons not to go jogging, viz. that it is hot and that it is raining. For a particular runner the combination of heat and rain may be less unpleasant than heat or rain alone so that the accrual is a weaker reason not to go running than the accruing reasons. And for another jogger the combination of heat and rain may be so pleasant that it is instead a reason to go jogging."*[67]

In this example, that it is hot (h) and that it is raining (r) are both reasons not to jog ($\neg\phi$), but the set composed of both reasons has no weight when it comes to not jogging – indeed, when taken together, the heat and rain count *in favour* of jogging. So, the weight of $(\{h, r\}, \neg\phi, w)$ is *smaller* than the weight of $(h, \neg\phi, w)$ and it is *smaller* than the weight of $(r, \neg\phi, w)$, in violation of Weak Monotonicity and Additivity.

There are also cases that appear to violate Additivity without violating Weak Monotonicity. To see this, consider the following example by Sher (2019):

> *"Suppose that I have a disease. My doctor proposes a treatment, and the question I am considering is, "Should I take the treatment?" … Suppose that r_1 is now the proposition that the treatment would prolong my life by at least 1 year, and r_2 is the proposition that the treatment would prolong my life by at least 2 years… The sum… of the weights of reasons r_1 and r_2 does not represent a meaningful quantity. This sum double counts the weight of the fact that the treatment will prolong my life by at least one year, as this fact is entailed by both r_1 and r_2."*[68]

This example satisfies Monotonicity, as the weight of the set composed of both reasons is at least as great as the weight of each of its elements – plausibly, $w(\{r_1, r_2\}) = w(r_2) > w(r_1)$, as the fact that the treatment will prolong his life by at least one year is entailed by the fact that it will prolong his life by at least two years, and thus contributes

[67] Prakken (2005), p. 86. Also appears in Nair (2021).

[68] Sher (2019), pp. 104-105. Also appears in Nair (2021). I replaced R_i with r_i to adapt the example to my notation.

no weight to the set $\{r_1, r_2\}$. But this is of course a violation of Additivity, because, if we are to avoid double-counting the following inequality must hold: $w(\{r_1, r_2\}) < w(r_2) + w(r_1)$. The example demonstrates that in some cases reasons "overlap" in the sense that summing their weights amounts to double counting, and thus, Additivity does not hold universally.

Both examples indicate that while in some cases Additivity appears like a natural explanation of the way individual reasons contribute to the weight of a set, in other cases reasons interact in ways that violate Additivity. Making sense of these phenomena requires a delineation of the scope in which Additivity holds, and an account of the ways reasons accrue when Additivity fails. Sher (2019) and Nair (2021) both develop sophisticated (but distinct) theories of weight that are designed to answer this challenge. They provide precise formulation of the phenomenon of "overlapping" reasons, and a precise way of measuring the extent to which reasons overlap. As a result, on both accounts the weight of a set of reasons is determined by two factors: (a) the weights of the individual reasons of which the set is composed and (b) the interactive properties involving those reasons that determine the degree to which they overlap. Reasons accrue additively precisely in the cases in which they are *independent* – when they do not possess the interactive properties that give rise to overlapping. The proposed pictures are therefore not Atomistic as the weight of a set of reasons is sensitive to factors beyond the weights of individual reasons, namely the interactive factors that determine the degree of "overlap" among them.

Both Sher and Nair initially help themselves to probabilistic tools to characterize weight of reasons. Sher understands the weight of a reason p to ϕ in terms of the expected value of ϕ-ing conditional on p. Nair characterizes the weight of a reason p to ϕ in terms of the degree to which p probabilistically confirms some proposition related to ϕ. The probabilistic nature of both accounts allows them to use different forms of probabilistic dependence to represent the interactive properties that underpin the phenomenon of reasons overlapping. Independence among reasons is defined by Sher in terms of relations among expected values conditional on them, and by Nair in terms of relations among degrees of conditional confirmation. Both authors then go on to stipulate a primitive weight function (not defined in terms of a probability function) and demonstrate how some purportedly intuitive conditions entail that these primitive weight functions have the same properties as those defined in probabilistic terms. Crucially, in both cases the primitive weight functions are stipulated in fully numerical terms and are not derived from a qualitative underlying structure.

Deriving a representation of weight with the properties argued for by Sher or Nair would require a significantly richer structure than the one explored here. Most importantly, such a structure would have to be able to express certain interactions among individual reasons within a set that give rise to a notion of reason "overlap". Such a derivation would be a combination of the project embarked on here – namely, that of deriving a numerical representation of the weight of reasons – and the project undertaken by Sher and Nair of characterizing the partially additive properties of such a representation. This would be a challenging and interesting way forward.[69]

In sum, the derived measurement of weight provided in the previous sections imposes no constraints on the relationship between weights of sets of reasons and the weights of their subsets. If Independence is false, and the notion of weight imposes some constraints on the relation between sets and subsets of reasons, then the representation of weight should be enhanced to accommodate, and preferably justify, such constraints. While accommodating some of the constraints discussed in the literature appears rather

[69] If Sher and Nair are right about weight having a probabilistic structure, then a possible way forward might be to explore the adaptation of representation theorems that derive probability functions from qualitative "likelier than" relations over propositional structures (e.g., those surveyed in Konek 2019), to the case of weight.

straightforward (like in the case of Monotonicity), the accommodation of the more nuanced constraints formulated by Sher and Nair would require considerable alteration of the representation derived here.

8. Conclusion

The weights of reasons are important. They figure saliently in the phenomenology of deliberation, and they play a central role in normative theorizing, especially in explaining the relation between reasons and facts about what one ought to do. Nonetheless, the notion of the weight of reasons is somewhat elusive – it is formulated metaphorically, and as is often the case with metaphors, it is not entirely clear how far the metaphor goes in describing it. It is not obvious which parts of the metaphor correspond to essential properties of the weight of reasons, and should therefore be taken seriously, and which parts are to be discarded as mere artifacts of the metaphoric representation.

This chapter is an attempt to make progress on this front. In particular, it offers a way of understanding the rich numerical structure of the weight metaphor as representing essential features of the weight of reasons. It demonstrates how numerical structures, and the expressive power they provide, can be understood as representations of intuitive numberless judgments about the weight of reasons. If the severity of outweighing relation among reasons satisfies some *prima fascia* intuitive constraints, then the weight of reasons is representable on a ratio scale. This implies that claims about some reason being twice as weighty as another (and a host of other numerical claims), can be understood in terms of claims about severity of outweighing between different pairs of reasons.

While the derived representation adequately represents many relations among weights of reasons, it does not accommodate some of the nuanced views concerning reason accrual and how individual reasons contribute to the weights of sets of reasons. In particular, it does not allow the expression of interactive factors that make reasons “overlap” and thus accrue non-additively. If these factors are essential features of the weights of reasons, then richer structures are required for the derivation of a measurement of weight that adequately reflects these interactions.

Chapter 4

The Norm of Assertion Is Not Epistemic

In this chapter I argue that the norm of assertion is not epistemic. I apply Schroeder's (2021) distinction between ϕ-ing right and ϕ-ing well to the case of assertion and consider the notion of *asserting well* – asserting permissibly with the right kind of motivation. The argument is composed of two parts. First, I demonstrate that if the norm of assertion is epistemic then asserting well requires one to be motivated to assert by facts about one's mental states. Second, I argue that asserting well cannot require this motivation, and therefore the norm of assertion is not epistemic.

1. Introduction

It is broadly accepted that assertion is governed by a norm. Some assertions are good, proper, or permissible, and others are bad, improper or impermissible *qua* assertions, and this is determined by a simple normative standard that assertions may satisfy or fail to satisfy, referred to as "the norm of assertion". While assertions may be subjected to a host of other more general norms, e.g., norms relating to politeness, rationality, or morality, the norm of assertion expresses a separate type of normativity that applies specifically to assertions *as such*. The norm of assertion is a condition of the following form:

One must: assert p only if C.

There is much disagreement about the content of C. Some take C to be a property of the asserted proposition independent of the asserter like truth, others take C to be a constraint on some relation between the asserter and the asserted proposition like knowledge. Epistemic norms of assertion are of the latter type, and take the following form:

One must: assert p only if one stands in epistemic relation R to p.

This class of epistemic norms is broad and is inhabited by many norms discussed in the literature. Substitute the epistemic relation R with knowledge and you get the knowledge norm of assertion.[1] Take R to be belief, and a belief norm is yielded.[2] Corresponding norms based on justified belief,[3] certainty, or any other epistemic relation, may be produced in the same way. This chapter argues that the norm of assertion is not epistemic, and thus targets any view that takes the norm of assertion to be an instance of the above schema.

The argument has two premises. First, that if the norm of assertion is epistemic then *asserting well* – a motivational standard to be elucidated below – requires the agent to assert for facts about her mental state. And second, that asserting well cannot involve such a requirement.

Those who are sceptical of the existence of a norm of assertion, and of the purported normative status of assertions *qua* assertions (versus their normative status as actions), may read my argument as having a conditional conclusion – if there is a norm of assertion, then it is not epistemic.[4]

2. Asserting Well

The first part of my argument relies on Schroeder's (2021) distinction between ϕ-ing right and ϕ-ing well. The original and best-known example of this schema is its moral case – the distinction between doing the morally right thing and acting with moral worth. Doing

[1] T. Williamson (1996).

[2] Hindriks (2007); Mandelkern & Dorst (2022).

[3] Douven (2006); Lackey (2007); Kvanvig (2009).

[4] Pagin (2011) considers the possibility that no norm governs assertion as such, but only argues for the weaker claim that no norm is constitutive of assertion.

the right thing amounts to acting in ways that comply with moral norms. Acting with moral worth entails doing the right thing but requires a further constraint on the agent's motivation for her (morally right) action. It requires her motivation to stand in a certain relation to the norm, or to some other normatively relevant features of her situation, in a way that makes her compliance with the norm non-accidental. For example, someone who donates to charity only to gain the admiration of their peers, does the right thing but fails to act with moral worth. There is something lacking in such a person's motivation and their doing the right thing is in an important sense accidental – had their peers not admired charitability they would not have acted charitably.

Schroeder observes that the distinction between doing the right thing and acting with moral worth is highly generalizable. One can make analogous distinctions between choosing the rational action and acting rationally, believing what one is justified in believing and believing justifiably, fearing what is appropriate to fear and fearing appropriately. Wherever there is a norm, Schroeder argues, we see such right-well pairs: ϕ-ing right which consists in *complying* with the norm, and ϕ-ing well which requires *following* the norm: complying non-accidentally, with the right kind of motivation. On this picture, if there is a norm of assertion, then it should yield a pair of standards: asserting right and asserting well.[5] The former is satisfied when one asserts in compliance with the norm and the latter is satisfied when one satisfies the additional anti-accidental motivational constraint.

My goal in this section is to argue that if the norm of assertion is epistemic, then asserting well – on the most plausible versions of this notion – requires the asserter to be motivated by facts about their own mental states.

[5] Lewis (2021) applies Schroeder's right/well distinction to assertion to argue for a "simple" knowledge norm (Williamson 1996) over an "express" knowledge norm (Turri 2011) for assertion.

There are roughly two types of views regarding the precise motivational constraint required for ϕ-ing well. On the first type of views, often termed *right reasons views*, ϕ-ing well requires, roughly speaking, being motivated by the reasons that make ϕ-ing right.[6] More precisely, views of this type require that the agent's motivational reasons for ϕ-ing match, in the appropriate sense, the normative reasons that make ϕ-ing right or permissible, i.e., the reasons for ϕ-ing. This matching requirement is meant to link the motivational and normative features of ϕ-ing in a way that rules out the accidentality and motivational inadequacy characteristic of failing to ϕ well. The thought is that if you ϕ for enough of the reasons that make doing so right or permissible, then your motivation is sufficiently sensitive to the normatively relevant features of your situation, and it is no mistake that you do the right or permissible thing.

Different right reasons views may specify different matching relations – different senses in which the agent's motivating reasons for ϕ-ing must "match" the normative reasons that make ϕ-ing right. For example, on a very strong view, ϕ-ing well requires that the agent be motivated by *all* the reasons for ϕ-ing. On a very weak view, it suffices to be motivated by *some* reason for ϕ-ing. I raise these examples, however, only to set them aside – neither of these extremes is plausible. The strong view makes ϕ-ing well impossible without being motivated by *all* reasons for ϕ-ing, no matter how negligible or redundant. In contrast, on the weak view, as long as the agent is motivated by *some* reason to ϕ – no matter how negligible or weak – they will count as ϕ-ing well. The most plausible right reasons views lie in between these extremes and require that the agent be motivated by a *sufficient* set of reasons for ϕ-ing. Sufficiency may be defined in different ways – e.g., in terms of the number, weight, or importance of reasons – but the overall rationale of such views is that to ϕ well, the agent must be motivated by considerations that are sufficiently normatively important; their motivation to ϕ must be sufficiently

[6] See Markovits (2010) and Schroeder (2021a).

normatively valanced.[7] If the agent fails to ϕ well, this is because they are missing something important – their motivation is too detached from the normatively relevant features of their situation.

Applied to the case of assertion, right reasons views entail that to assert well, one must be motivated by the reasons that make asserting permissible.[8] But, at least on the more plausible sufficiency-based views, one need not be motivated by *all* such reasons, rather, it suffices that the set of motivating reasons be *sufficient* in the relevant sense. What type of facts are the reasons that make asserting permissible? Which sets of reasons qualify as sufficient for asserting well? If the norm of assertion is epistemic, then the fact that the asserter bears the required epistemic relation to the asserted proposition – let us call this the *satisfaction fact* – is a reason that makes asserting permissible. Indeed, the satisfaction fact is in an important sense the best, or most important reason, as it determines the permissibility of an assertion. However, depending on how the notion of sufficiency is spelled out, and which type of facts are reasons for assertion, it may be possible to assert well without being motivated by the satisfaction fact – without being motivated by the fact that one bears the required epistemic relation to the asserted proposition. Perhaps, being motivated by related facts regarding the asserter's epistemic state, or their mental state more generally, could be sufficient for asserting well.

However, on any plausible precisification of sufficiency, it is hard to see how a set of reasons that includes *no* fact about the asserter's mental states – and therefore, no fact

[7] For example, Schroeder (2021a) argues for a notion of sufficiency on which a set or reasons is sufficient for ϕ-ing if it outweighs the reasons against ϕ-ing.

[8] One might worry that right reasons views are inconsistent with an epistemic norm of assertion altogether because they are committed to the norm of assertion being determined by the balance of reasons, and whether an asserter bears epistemic relation R to p is not determined in this way. However, as Schroeder (2021: 219-220, 226) notes, norms that are not naturally formulated in terms of reasons may nonetheless be determined by the balance of reasons. In the case of assertion, as long as the fact that one doesn't bear R to p is always a better reason against asserting p than any reason in favor of doing so, epistemic norms for assertion may be determined by the balance of reasons.

about their epistemic states – could count as sufficient, if the norm of assertion is epistemic. Recall, that the notion of sufficiency is meant to prevent the agent's motivation from veering too far from the normatively relevant features of their situation. A set of motivating reasons to ϕ is sufficient only if it includes enough of what is of normative concern when it comes to ϕ-ing. If assertion is governed by an epistemic norm, then the normatively relevant features of an asserter's situation are epistemic, or at least mental – a description of such a situation that makes no reference to the asserter's mental states must be missing something crucial about the normative forces at play.[9] Therefore, a set of motivating reasons for assertion that includes *no* fact about the asserter's mental states suffers from the precise defect that the notions of sufficiency, and more generally of ϕ-ing well, are meant to prohibit – it is too divorced from the normatively relevant features of the situation vis-à-vis assertion. Therefore, asserting well – in the *right reasons views'* sense – requires being motivated by facts about one's own mental states, if the norm of assertion is epistemic.[10]

On the second type of views regarding the motivational constraint required for ϕ-ing well – which I will term the *right views* – ϕ-ing well requires being motivated by the very fact that ϕ-ing is right, or permissible.[11] Applied to assertion, such views entail that

[9] This claim is consistent with there being *some* non-mental normatively relevant features when it comes to assertion. It only states that some mental features are normatively indispensable – any description of the asserter's situation that omits *all* mental facts would be problematically partial. For example, on a knowledge norm for assertion, truth is surely a normatively relevant feature – it is part of what it takes for an assertion to be permissible. However, a description of the asserter's situation that refers to the truth of the asserted proposition but omits all details about their mental states – e.g., that they believe the asserted proposition, that they justifiably believe it, that they have high credence in it – would be missing something crucial. In turn, an asserter motivated only by the truth of their assertion, but not by any fact about their epistemic or mental states, could not be asserting well if assertion is governed by a knowledge norm.

[10] My argument here exploits the relation between the content of norms and normative reasons and may be generalized and perhaps applied in other contexts. In a more abstract from, the argument relies on the fact that a norm for ϕ-ing typically picks out a neighborhood of considerations that are normatively relevant vis-à-vis ϕ-ing, and therefore places constraints on what it takes to ϕ well. If these motivational constraints are implausible, then this can count against the proposed norm for ϕ-ing.

[11] See Johnson King (2020). Sliwa (2016) requires in addition that one know that ϕ-ing is right.

asserting well requires being motivated by the very fact that doing so is permissible. It might be initially tempting to identify the fact that asserting p is permissible with the fact that asserting p satisfies the permissibility condition cited by the norm of assertion. On this line of reasoning, if the norm of assertion is epistemic, then asserting well would require being motivated by the fact that one bears the required epistemic relation to the asserted proposition. Therefore, asserting well requires being motivated by facts about the asserter's mental states also on right views of ϕ-ing well.

However, this would be too quick. Even if the expressions "asserting p is permissible" and "the asserter bears R to p" are extensionally equivalent, indeed, even if they are intensionally equivalent, it is clearly possible to be motivated by the fact that asserting p is permissible without being motivated by the fact that one bears R to p. Think of an agent who somehow knows – perhaps via reliable testimony – that asserting p is permissible but has no idea that they bear R to p, or that permissibility for assertion is determined this way. Such an agent could easily be motivated to assert p by the fact that doing so is permissible without being motivated to do so by the fact that they bear R to p.[12]

So, the well standard for assertion on the *right views*, does not require being motivated by facts about the asserter's mental states. Indeed, the emerging well standard is quite independent of the content of the norm of assertion. All that asserting well requires is that the agent be motivated by the "opaque" normative fact that asserting is permissible, and this is consistent with the content of the norm of assertion (and the

[12] If the fact that asserting p is permissible and the fact that one bears R to p are the same fact, then the objects of motivation must be more fine-grained than facts. Perhaps facts under certain descriptions (see Johnson King 2022). One may adopt this view and adjust the discussion in the text accordingly by replacing facts with facts under certain descriptions as the objects of motivation.

related normative considerations) being absent from the asserter's motivation, and indeed from their knowledge or awareness.

Therefore, if *right views* are correct, the content of the norm of assertion places *no* constraints on the type of motivation required for asserting well, and thus my argument concerning the consequences of an epistemic norm of assertion does not go through. While my argument can be read primarily as conditional on the right reasons views, I will use the remainder of this section to raise two doubts about the plausibility of the right views' well standards for assertion and other non-moral domains.[13]

First, the right views' condition for ϕ-ing well, namely that the agent be motivated by the very fact that ϕ-ing is right, seems contrived and unreasonably demanding in non-moral cases. Consider for example, what it would take to fear well on such a view: one would have to be motivated to fear by the fact that fearing is right, or appropriate, in one's situation. But it would be extremely odd if this high-level normative fact – about the appropriateness of fearing in one's situation – figured in one's motivation to fear. A normal person would be motivated to fear by facts like the proximity of a bear, or the raging of a storm, not by normative facts about when it is appropriate to fear.[14] In contrast, right reasons views do a much better job of vindicating this motivational pattern – the proximity of a bear and the raging of a storm are reasons for fearing, they make it appropriate to fear, and are therefore the type of fact we should expect to motivate an agent that fears well on such views.

Secondly, and perhaps more importantly, the detachment between the content of the norm and the motivational condition for ϕ-ing well, makes right views implausible,

[13] Proponents of right views, as far as I am aware, only argue for them in the moral case (see Sliwa 2016 and Johnson King 2020). The doubts raised below are targeted at the generalization of these views to non-moral domains such as assertion.

[14] See Schroeder (2021a), p. 217 for this type of criticism of the *right views*.

at least in non-moral domains. To see this, consider the following example which applies a right views conception of ϕ-ing well to the case of chess – the rules of chess make it the case that some moves are right, and some are wrong, and therefore the right/well distinction is applicable:

> **Phill** is playing chess online. While Phill knows nothing about chess, his friend Deborah is a world champion and is therefore an extremely reliable source when it comes to the game. Deborah tells Phill (correctly) that moving his knight to d4 would be the right move in his situation. Having no idea why this move is right, Phill goes ahead and moves his knight to d4. He is motivated to do so by the fact that it is the right move.

If *right views* are correct, then Phill plays chess well (in the technical sense of this term). While the type of standard that Phill satisfies here might be normatively interesting in some ways, it clearly leaves much to be desired. Phill is making the right move because it is right, but he does not engage with the normatively important features of his situation. He is oblivious to the myriad of considerations that count in favour of making the move, indeed he is unaware of the very rules of the game. If we want a well standard that can account for the normative importance of these facts, and for the motivational shortcomings of Phill, then we must reject right views. In contrast, right reasons views yield well standards that do not suffer from these shortcomings – on no such view would Phill be considered playing chess well. To play well according to such views, there would have to be a stronger link between his motivation and the content of the norms governing his situation, namely the rules of chess.[15]

[15] The shortcomings of the right views elicited by this chess example apply also in the case of assertion. Consider the following example:

> **Amos** is deliberating whether to assert some proposition p. While Amos has no idea whether doing so would be permissible or what the correct norm of assertion is, his friend Danny is an expert on norms of assertion and is therefore a reliable source when it comes to such matters. Danny tells

In sum, on right reasons views, if the norm of assertion is epistemic, then asserting well requires being motivated by facts about the asserter's mental states. While this conclusion doesn't follow on the right views' notion of ϕ-ing well, the generalization of such views beyond the moral case is less plausible.

3. A Faulty Well Standard

The previous section established that on an epistemic norm of assertion, asserting well requires asserting for what I will term *mental motivations* – being motivated to assert by facts about the agent's own mental states. In this section I provide two arguments for the claim that asserting well should *not* require such motivations. It therefore follows that the norm of assertion is not epistemic.

The first argument relies on the claim that failing to assert well, like failing to act or believe well, should be normatively problematic – we should be able to detect a flaw in the asserter's motivation and an accidentality in her asserting rightly. However, upon reflection, there seems to be nothing wrong with asserters who violate the well standard yielded by epistemic norms for assertion.

To see this, pick your favourite epistemic norm for assertion, and suppose that there's a rail strike today and that you bear the required epistemic relation R to that proposition. You see your neighbour Harvey head out to the train station and think to yourself: 'poor Harvey, there's a rail strike today, so he'll end up walking to the train station in vain. I should tell him about it to save him the frustration'. You then go on to tell Harvey that there's a rail strike today. In doing so, let us assume, you are not

Amos (correctly) that asserting p is permissible for him according to the correct norm of assertion. Amos then goes on to assert p. He is motivated to do so by the fact that it is permissible.
Amos's case suffers from the same problems as Phill's. They exhibit analogous motivational vices.

motivated by the fact that you bear the epistemic relation *R* to the proposition that there's a rail strike today. Indeed, you are not motivated by any fact regarding your epistemic state. Rather, you are motivated simply by the fact that there's a rail strike today (and by the other relevant facts about Harvey's preferences, the relevance of the rail strike to them, etc.).[16]

Is there anything wrong about your motivation to assert? Are you in any sense asserting for the wrong reasons? Are you in any way similar to the person who gives to charity to gain admiration? It is, I think, quite clear that the answers to these questions are negative. Your thought leading to the assertion seems like an ordinary and sound piece of reasoning. It is sensitive to features that are clearly relevant to the question of whether to assert that there's a rail strike, and it doesn't appear to miss anything crucial. Your assertion does not suffer from the paradigmatic problems of violating a well standard: there seems to be nothing amiss about your motivation and nothing objectionably accidental about your assertion's conforming to any normatively relevant standard.

The well standard yielded by an epistemic norm of assertion doesn't behave like it should – it doesn't appear to pick out the normative property characteristic of such standards, and violations of it seem normatively innocuous. If the well standard for assertion must (a) stand in the required relation to the norm of assertion (as characterized in the previous section) and (b) have the normative properties characteristic of well standards (proper motivation and non-accidentality), then the norm of assertion cannot be epistemic.

[16] Surely the fact that you bear *R* to *p* is part of the causal story *explaining* your assertion. But it need not be part of your motivation to assert. Indeed, as the typicality of the case suggests, such facts are often *not* part of one's motivation.

The second argument relies on the thought that when one asserts well, one should not be asserting for the wrong reasons – for facts that do not amount to considerations in favour of assertion. This argument draws on Schroeder's (2008) distinction between objective and subjective reasons. An objective reason for an agent to ϕ is a proposition that counts in favor of her ϕ-ing, regardless of whether she believes the proposition. It is the type of reason that distinguishes Rikki from Bobby:[17]

> **Rikki** needs to go to work today. She normally prefers to commute by train rather than cycle. Unbeknownst to her, there's a rail strike today.
>
> **Bobby** is working from home today. He has the same commute preferences as Rikki, and like her, he is ignorant of the rail strike.

While there's a reason for Rikki to cycle to the office today, there is no such reason for Bobby (as he's working from home). The sense of 'reason' that makes this statement true, is the sense alluded to by the term 'objective reason'.

A subjective reason for an agent to ϕ is a proposition that she takes to count in favor of her ϕ-ing, and if it is true, it is an objective reason, i.e., it counts in favour of her ϕ-ing.[18] It is the type of reason that distinguishes Rikki (above) from Frida:

> **Frida** also needs to go to work. She has the same commute preferences as Rikki, but unlike Rikki, Frida knows about the rail strike.

[17] Rikki, Bobby and Frida (below) correspond to Schroeder's (2008) Ronnie, Bradley and Freddie.

[18] Schroeder (2009) points out that this conditional definition may break down in cases where the agent falsely believes p, and in the closest possible world(s) in which it is true, it is not an objective reason to ϕ because of some change in the agent's circumstances. For example, if Betty falsely believes there's a rail strike today, the proposition that there's a rail strike today, may still be a subjective reason for her to cycle to work even though in the closest possible world in which there is a rail strike, Betty is on leave. To cope with this complication, Schroeder suggests an alternative definition according to which a proposition is a subjective reason for an agent to ϕ if it has the property, if it is true, of making it the case that the agent has an objective reason to ϕ (See Schroeder 2009, pp. 230-233). For simplicity, I will stick to the conditional definition in the text as it will do for my purposes here.

While there are objective reasons for both Frida and Rikki to cycle to work, Frida *has* a reason to do so in a sense that Rikki does not. It is this sense of 'reason' that is alluded to by the term 'subjective reason'.[19]

Back to assertion. While unlocking her bike, Frida sees her neighbour Harvey leaving toward the train station and tells him that there's a rail strike today (p). Let us assume that Frida bears the epistemic relation to the asserted proposition required by your favourite epistemic norm for assertion. Which of the following is an objective reason for Frida to assert?

(1) That she bears R to the proposition that there's a rail strike.

(2) That there's a rail strike.

To answer this question, it would be useful to consider two counterfactual cases, one in which (1) is true and (2) is false and one with the opposite truth assignment, and see how Frida's reasons vary.[20] However, in doing so we must distinguish between factive and non-factive epistemic relations. When the epistemic relation R is non-factive, (1) and (2) are independent and we may consider both counterfactuals. But for a factive epistemic relation R, (1) entails (2) and thus only the second counterfactual is possible.

First, assume R is non-factive, and consider the case in which Frida bears R to the proposition that there's a strike, while in fact there is no strike. Does counterfactual-Frida have an objective reason to assert that there's a rail strike? It appears that she does not. Telling Harvey that there's a rail strike would not promote any of the goals of doing so – it would not inform or help him in any way. Indeed, the considerations in favour of actual-Frida asserting p don't appear to apply in this counterfactual case. At the very least,

[19] While the objective reason relation is factive – that p is an objective reason for someone to ϕ entails p – the subjective reason relation is not: the agent might have a false subjective reason to ϕ. If Frida were wrong about the strike, that there is a strike would still be a subjective reason for her to cycle.

[20] See Enoch (2010, 2015) for cases involving reasons for belief and action respectively.

counterfactual-Frida appears to have lost much of the objective reason that actual-Frida has for asserting, and the fact that counterfactual-Frida bears R to p does not seem to prevent this. While counterfactual-Frida might have a *subjective* reason to assert – if she believes that p – unlike actual-Frida, she lacks an objective reason to do so. This counterfactual test demonstrates that Frida's objective reasons to assert p vary with the truth of p even when her epistemic relation to p is held fixed. This suggests that while (2) *is* a reason for actual Frida to assert, (1) is not.

The second counterfactual test can be run on both factive and non-factive epistemic relations. Consider the case in which there is in fact a rail strike, but Frida fails to bear the required epistemic relation R to this proposition. Does she have an objective reason to assert? While she might fail to have a *subjective* reason (if she doesn't believe that p), she arguably still has an objective reason. That there is in fact a rail strike, is still a consideration in favour of her asserting as much – it would promote the goals of the assertion, it would inform and aid Harvey (the same goals of actual-Frida's assertion). While Frida might lack access to this consideration and therefore might be unable to act on it, it is a consideration, nonetheless.[21] This counterfactual test demonstrates that Frida's objective reasons to assert p do not vary with her epistemic relation to p when the truth of p is held fixed. This too suggests that p, not Frida's epistemic relation to p, is her objective reason to assert.

The counterfactual tests demonstrate that while the content of an assertion *is* an objective reason to assert, the agent's epistemic relation to its content is not. So, if the norm of assertion is epistemic, then asserting well requires being motivated by facts that are not objective reasons to assert.

[21] Compare counterfactual-Frida's reason to assert to Rikki's reason to cycle.

However, perhaps these facts that are not objective reasons to assert – like the agent's epistemic relation to the content of the assertion – are nonetheless subjective reasons to do so? Perhaps epistemic norms of assertion give rise to a *subjective* well standard requiring some matching between the agent's motivation and their *subjective* reasons? After all, there is a sense in which our evaluation of the counterfactual versions of Frida is sensitive to her subjective reasons. In the first counterfactual case, Frida would be doing the best she could, if she told Harvey that there's a rail strike – assuming that she believed as much – even though there is in fact no strike. Her assertion would be an apt response to her subjective reasons – to what she believes is in fact a consideration in favour of asserting, and the unfortunate fact that she's wrong does not undermine this aptness. Indeed, we'd be critical of Frida had she not told Harvey as much, as this would be a failure to act on her subjective reasons, and from her perspective, she'd be failing to do something she has reason to do – she'd be withholding important information from Harvey.

Similarly, in the second counterfactual case, if Frida fails to believe that there's a rail strike (when in fact there is one), then she lacks a subjective reason to assert, and therefore, not asserting would be an appropriate response to her subjective reasons (or lack thereof). Indeed, we'd be critical of her if she were to tell Harvey that there's a strike as she would take herself to be lying (or at least asserting a proposition she did not believe) and would not be aptly responding to her lack of subjective reasons to assert. So, perhaps the well standard for epistemic norms of assertion tracks this pattern of evaluation – requiring the asserter to be motivated by (enough of) their subjective reasons?

There very well may be a subjective well standard for assertion, requiring the agent's assertion to be motivated by their subjective reasons to assert. However, such a standard would not require the agent to be motivated by the fact that they bear some epistemic relation to a proposition, as this fact cannot be a subjective reason to assert. Recall that a subjective reason is a proposition that the agent believes, and if it is true, it is an objective reason to assert. But the above discussion of the counterfactual cases demonstrated that the proposition "the agent bears the required epistemic relation to the asserted proposition" is not an objective reason to assert, even when it is true. Therefore, this proposition cannot be a subjective reason to assert even if the agent believes it, as its truth doesn't make it an objective reason.

So, the fact that the agent bears some epistemic relation to the asserted proposition, and other facts about their mental states, are generally neither subjective nor objective reasons to assert. And thus, if the norm of assertion is epistemic, then asserting well entails asserting for *bad reasons* – propositions (or facts) that are neither objective nor subjective reasons to assert. But this cannot be the case, since asserting well consists in asserting for the right kind of motivation. Therefore, the norm of assertion cannot be epistemic.

4. Towards a Positive Account

If the argument in the previous sections is sound, then the norm of assertion cannot be epistemic, as it gives rise to an erroneous notion of asserting well. While the main aim of this chapter is negative, the argument above suggests a new criterion for assessing norms of assertion – namely, whether they generate a well-behaved well standard – which may direct us toward a positive account for the norm of assertion. In this section I will take some first steps toward such an account.

A natural place to start is a simple truth norm:

One must: assert p only if p.

What type of motivation would be required for asserting well on such a norm? The *satisfaction fact* for the truth norm is that p is true, but as discussed in section 2, other related facts may amount to reasons for assertion and may figure in the motivation of an agent who asserts well. There is more than one way of delineating the type of facts that make assertion permissible given a truth norm, but let us assume for now that they are evidential – the type of consideration that makes an assertion permissible is evidence for the asserted proposition. Therefore, while asserting a proposition well does not require being motivated by its truth, it does require being motivated by sufficient evidence for it. The satisfaction fact – that p is true – is, in an important sense, the best reason to assert as it is the best evidence for p, but one may assert well even without p being among their motivating reasons, as long as one is motivated by enough evidence for p.

This notion of asserting well is vindicated by the examples in the previous section – telling Harvey that there's a train strike for the reason that there's a train strike appeared like a typical and well-motivated assertion. Further, the counterfactual tests suggested that while the asserter's epistemic relation to p is typically *not* an objective reason to assert p, the fact that p, *is* such a reason. So, the considered truth norm is not susceptible to the arguments against the epistemic norms, as it gives rise to a sound well standard.

The well standard yielded by the truth norm has two other desirable features that might make the truth norm appealing. First, it amounts to an anti-Bullshit condition in the Frankfurtian (2005) sense, and therefore tracks an independently plausible normative standard.[22] Frankfurt characterizes Bullshit as the practice of asserting "without any

[22] Lewis (2021) makes the link between failures to assert well and Bullshit. However, he operates with an epistemic notion of asserting well – asserting p for the reason that one knows p. This is precisely the type of motivation argued against in the previous section. It also does not contrast precisely with Bullshit – it does not ensure that the asserter is motivated by the truth of what they say, rather by their epistemic relation to what they say. When the epistemic relation is factive, this will entail truth, but as far as Bullshit is concerned, their epistemic relation is beside the point, what matters is whether they are motivated by the

regard for how things really are",[23] or in a manner "disconnected from a concern with the truth".[24] While both an ordinary asserter and a liar are concerned with the truth value of what they assert (the former trying to report the truth, the latter trying to conceal it), the Bullshitter is characterized precisely by having no such concern: they may be motivated to assert by a variety of factors, but the truth of their assertion is not among them. Bullshithood is therefore a motivational property, it depends not on the content of the assertion – Bullshit need not be false – but on the asserter's motivation to assert.[25] The relation between asserting truly and Bullshitting is very similar to the relation between doing the right thing and failing to act with moral worth. While a Bullshit assertion need not be false, when it is true, it is, in a normatively important sense, accidentally true – the asserter wasn't aiming at the truth, they just happened to hit it. And, like failing to act with moral worth, there is something amiss with the Bullshitter's motivation. So, Bullshitting has the characteristic properties of failing a well standard for assertion, where the corresponding right standard is truth. Furthermore, the well standard yielded by the truth norm is the precise negation of Bullshit – it requires that the asserter be motivated by considerations relating to the truth of what they assert, namely by evidence for it. This counters the precise detachment between motivation and truth that is constitutive of

truth. I therefore agree with Lewis that the well standard for assertion is an anti-Bullshit condition, but I disagree with him about what it takes to assert well.

[23] Frankfurt (2005), p. 30.

[24] Ibid, p. 40.

[25] In Frankfurt's terms, whether an assertion is Bullshit depends not on its content but on how it was produced:

> "For the essence of bullshit is not that it is false but that it is phony. In order to appreciate this distinction, one must recognize that a fake or a phony need not be in any respect (apart from authenticity itself) inferior to the real thing. What is not genuine need not also be defective in some other way. It may be, after all, an exact copy. What is wrong with a counterfeit is not what it is like, but how it was made. This points to a similar and fundamental aspect of the essential nature of bullshit: although it is produced without concern with the truth, it need not be false. The bullshitter is faking things. But this does not mean that he necessarily gets them wrong." (pp. 47-8)

Bullshit. So, the well standard yielded by the truth norm maps onto an independently motivated normative standard – that of refraining from Bullshit.

The second desirable feature of the considered truth norm is that it may give rise to debunking arguments against epistemic norms of assertion, explaining why such norms are misleadingly appealing. To see how this could work, let us first consider the case of an agent who is motivated to assert p by the satisfaction fact, that is, by the fact that p is true. Let us call this stronger motivational standard – being motivated by the satisfaction fact – the *very well* standard, and let us therefore say that such an agent asserts *very well*.[26] Now, asserting very well could entail bearing some epistemic relation to p, and there are two routes to such an entailment. First, if treating a proposition as a reason *conceptually entails* bearing some epistemic relation to it then asserting p very well entails bearing that relation to p. For example, assume that part of what it takes to treat p as a reason to ϕ is to believe that p. Then whenever an agent asserts p very well, they must also believe that p, because asserting p very well requires them to assert for the reason that p, and treating p as a reason involves believing p. Similarly, if treating p as a reason must involve knowing that p then knowing what one asserts will be a by-product of asserting very well. More generally, for any epistemic relation R, if bearing R to p is part of what it takes to treat p as a reason, then bearing R to p will be a by-product of asserting p very well.

The second route by which asserting p very well may entail bearing certain epistemic relations to p, is normative. If, to treat a proposition as a reason one must – as a normative matter – bear some epistemic relation to it, then whenever one asserts p very well, one will be required by this normative constraint to bear the relevant epistemic

[26] As mentioned above, such an agent would be in some sense motivated by the "best" reason to assert p, as p is the best evidence for p.

relation to p. For example, assume one accepts a knowledge norm for practical reasoning of the following sort:[27]

Treat p as a reason for action only if you know that p.

Given such a norm, whenever one asserts p very well one would be required to know p.[28] Other norms for practical reasoning – e.g., ones that require belief, justified belief, sufficiently high credence, etc. – would yield other epistemic relations as normative by-products of asserting very well. In general, for any epistemic relation R, if one ought to treat p as a reason only if one bears R to p, then whenever one asserts p very well one ought to bear R to p.[29]

If asserting very well is a genuine normative standard then these two ways in which epistemic relations might reemerge as conceptual or normative by-products of asserting very well, could provide potential grounds for debunking arguments against the epistemic norms for assertion. For, if asserting p very well entails bearing some epistemic relation to p, that could explain why epistemic norms of assertion are tempting – they correlate very well with a genuine normative standard for assertion, namely with the very well standard. However, as the negative argument in the previous sections demonstrated, identifying the norm with these entailed conditions would be a mistake.[30]

[27] Hawthorne & Stanley (2008) argue for such a norm. See also Hyman (1999), T. Williamson (2017). For a decision-theoretic explication see Goldschmidt (2024).

[28] Whiting (2013) argues for a truth norm of assertion from a knowledge norm for practical reasoning.

[29] This would also explain why the epistemic standards for assertion and practical reasoning are identical – e.g., both require knowledge – since whenever one asserts very well one treats the asserted proposition as a reason and is thus required to bear the epistemic relation required for practical reasoning to the asserted proposition. See J. Brown (2019) who fails to find a satisfying explanation for the purported identity of these standards.

[30] One might then argue that the epistemic norms were meant as well standards to begin with. However, this cannot be right because bearing epistemic relation R to p places no constraints on the asserter's motivation to assert. A condition that places no constraints on the motivation to ϕ cannot be a standard of ϕ-ing well.

However, even if one is not convinced that the very well standard is of any genuine normative importance, similar, though admittedly weaker, debunking arguments could still be produced. To see this, notice that given a truth norm, asserting p well requires being motivated by facts that amount to sufficiently strong evidence for p, and therefore treating such facts as reasons. The two modes of entailment described above could then apply to *these* facts, resulting in conceptually or normatively entailed epistemic relations – e.g., belief or knowledge – to a set of propositions that is sufficient evidence for p. So, whenever an agent asserts p well, they would, as a conceptual or normative matter, have to bear certain epistemic relations to sufficient evidence for p. Now, while believing or knowing a set of propositions that amounts to sufficient evidence for p, and believing or knowing p, are not the same, they are also not wildly different. Therefore, it is possible that epistemic norms for assertion are driven by intuitions about entailed epistemic relations towards evidence for the asserted proposition.

A second debunking argument that does not depend on a notion of asserting very well, turns on the type of sufficiency required for asserting well. If for a set of propositions to be sufficient for asserting p it must also be sufficient for bearing some epistemic relation to p, then whenever an agent asserts p well, they would be in a position to bear that epistemic relation to p. For example, suppose that the sufficiency relations for assertion and belief are sufficiently similar such that whenever a set of propositions is sufficient for asserting p it is also sufficient for believing p. Then whenever one asserts p well, one would be motivated by a set of reasons that would be sufficient for believing p – if one were to believe p on the basis of those same reasons, one would be believing p well. If this is on the right track, to assert well one must be in a position to believe well. It is easy to see how this correlation between asserting well and being in a position to believe well could lead one to think that one must believe, or even believe well, what one asserts, and that assertion is governed by a corresponding belief norm.[31] Similar arguments could be produced for other epistemic relations (besides belief) if the relevant sufficiency relation is similar enough to the assertion one.

Far from a conclusive argument for a truth norm for assertion, this section provides a preliminary demonstration of how one may go about searching for a norm of assertion that yields a well-behaved well standard. It is also a demonstration of the manner in which a satisfactory norm of assertion might provide debunking arguments against some epistemic norms.

5. Conclusion

The norm of assertion cannot be epistemic because this would entail an erroneous standard of asserting well – a standard that it is both normatively unproblematic to violate, and whose satisfaction entails a motivational vice. While the conclusion of this chapter is negative, the argument suggests a criterion for assessing norms of assertion – adequate norms must give rise to well-behaved notions of asserting well. While epistemic norms fail this criterion, other norms might not.

[31] If Schroeder (2021a) is right, and knowing p is believing p well, then asserting p well entails being in a position to know p, which could provide grounds for a debunking argument against the knowledge norm for assertion.

Chapter 5

Taste Predicates and Clouds of Contexts

Contextualist theories for taste predicates like “tasty” and “fun” take such predicates to be perspective-relative and require the context to fix the relevant perspective for their evaluation. On this picture, saying that something is tasty is saying that something is tasty to some contextually-determined individual or group. However, dialogues involving disagreement about taste generate problems for such theories as there seems to be no way of fixing the perspectives in question while respecting linguistic intuitions. In particular, the contextualist appears unable to account for the intuition that such dialogues involve faultless disagreement – genuine disagreement in which no party is at fault. These problems have been used to propel arguments against contextualism and for alternative theories like relativism.

In this chapter I suggest a solution for the contextualist. I propose adopting von Fintel and Gillies’ (2011) theory of “Clouds of Admissible Contexts” to the case of taste predicates. Their theory is tailored to solve a problem for the contextualist in relation to epistemic modals but – I will argue – is also able to solve their problems relating to taste predicates due to the similarity between these two linguistic phenomena.

The structure of the chapter is as follows. In section 1 I will present contextualist semantics for taste predicates, in section 2 I will demonstrate how some dialogues involving disagreement pose a problem for the contextualist, and in section 3 I will introduce von Fintel and Gillies’ theory, apply it to the case of taste predicates and briefly motivate this application. Finally, in section 4 I will conclude.

1. Contextualist Semantics for Taste Predicates

It is widely thought that the meaning of taste predicates such as "tasty" and "fun", is sensitive to someone's, or some group's perspective. For example, the meaning of sentences like:

(1) Lemon pie is tasty.

is sensitive to the perspective of someone or some group for whom lemon pie might be tasty. So, the truth value of (1) depends not only on facts regarding lemon pie, but also on facts about its relation to some implicit subject or set of subjects.

This thought – that perspective plays a central role in the semantics of taste predicates – is motivated by the combination of two facts. First, the prevalence of disagreement regarding matters of taste, and second, the typical lack of epistemic humility and cautiousness in judgments about taste. If the truth value of sentences involving taste predicates were perspective-neutral – like the truth value of sentences such as "lemon pie is Gluten free" – then the degree of disagreement regarding matters of taste would suggest that many of us are pretty poor judges of such matters and should have low confidence in our judgments about them. However, we do not hesitate to judge things as tasty and funny, and we do not appear to take disagreement to warrant lower confidence in such judgments. Making the semantics of taste predicates perspective-relative (in some way or another) allows vindicating these ubiquitous linguistic and epistemic practices.[1]

How exactly does perspective figure in the semantics of taste predicates? According to contextualism, perspective plays a role in fixing the content of sentences

[1] I take this point from MacFarlane (2007).

with taste predicates.[2] On this view, taste predicates are equipped with a covert perspective argument that is determined by the context. "Tasty" and "fun" thus always mean "tasty to ___" and "fun for ___", where the blanks are filled with some perspective bearer (or bearers) determined by the context. Contextualist theories typically take the speaker to be the default perspective bearer, and thus (1) uttered by Rachel would typically express the proposition that lemon pie is tasty to Rachel, whereas (1) uttered by Lupo would typically express the proposition that lemon pie is tasty to Lupo. In such cases taste predicates behave similarly to indexicals.[3] Consider:

(2) I live in London.

Since the reference of the indexical "I" depends on the speaker, the context determines the content of (2). Thus, when uttered by Rachel, (2) expresses the proposition that Rachel lives in London and when uttered by Lupo, (2) expresses the proposition that Lupo lives in London.

However, in some cases, and on some contextualist theories, the content of taste sentences may be determined by other perspectives besides the speaker's. Consider for example Rachel uttering (3) while watching her cat enthusiastically devour a bowl of a new brand of cat food:

(3) This cat food must be tasty.[4]

Or consider for example, Lupo uttering (1) after his mother tells him she wants to bake something for him and Rachel and asks what they like.[5]

[2] See Lasersohn (2005), sections 1-2 for a detailed account (that he argues against). See Glanzberg (2007) for a defence of contextualism.

[3] Indeed, on some contextualist theories, taste predicates *are* indexicals. See Schaffer's (2011) *Indexicalism*.

[4] This example is an altered version of an example discussed in Stephenson (2007), p. 498.

[5] This example is an altered version of an example discussed in Schaffer (2011), p. 217.

A plausible contextualist story for these examples is that given the contexts, (3) expresses the proposition that the cat food must be tasty *to the cat*, and (1) expresses the proposition that lemon pie is tasty *to Lupo and Rachel*. Contextualism may thus be rather flexible in the way it allows contexts to pick out the relevant perspectives for the evaluation of taste predicates, and the relation between contexts and the contents of taste sentences may ultimately be quite intricate.

Before I move on to explore some problems for contextualism, it would be useful to present briefly a prominent theoretical alternative – relativism. For the purpose of contrasting the theories, it would be helpful to recall Kaplan's (1989) semantic framework. According to Kaplan, the content of a sentence is fixed by a context, and this content – a proposition – receives a truth value relative to an index. For example, a context in which Lupo utters (2) fixes the content of the sentence – namely, the proposition that Lupo lives in London – and the index from which the sentence is evaluated – e.g., the actual world, now – determines its truth value.

Given this framework, let us return to the semantics of taste predicates. While contextualism inserts perspective in the first Kaplanian stage – that of content fixing – relativist theories introduce perspective only at the index stage relative to which truth value is evaluated and after content is fixed. The relativist introduces perspective at this stage by supplementing the traditional index with an additional coordinate, that of a judge. Therefore, the truth value of a sentence on this view is no longer relative to merely a world and time of evaluation, but to a judge as well. A proposition may thus be true as judged by me and false as judged by you, in a manner analogous to it being true in one possible world and false in another. The relativist takes the judge coordinate to take the value of the *assessor* which would typically be the speaker but need not be.[6]

[6] See Lasersohn (2005) and Stephenson (2007).

For the contextualist when Rachel and Lupo each say (1), they are expressing two different propositions. In contrast, for the relativist, they are expressing the same perspective-neutral proposition, but the index relative to which its truth value is evaluated differs between their assertions. Specifically, the judge coordinate of the index receives different values for the two assertions because it (typically) takes the value of the speaker.

2. The Problem of (Dis)agreement

The central challenge faced by contextualist semantics and the main motivation for adopting relativism is the way the former deals with cases of disagreement on matters of taste. Specifically, it appears that contextualism cannot accommodate the fact that disagreements regarding taste appear to be both genuine and faultless. Consider the following dialogue:

(4) *Rachel*: Lemon pie is tasty.
Lupo: No, lemon pie is not tasty.[7]

Two seemingly contradicting intuitions emerge from this dialogue. On the one hand, Rachel and Lupo seem to be genuinely disagreeing, and on the other hand, neither seem to be at any fault – neither seems to be mistaken in any way.[8] While contextualism can clearly accommodate the second intuition, it appears not to do justice with the first. On a simple contextualist view, Rachel's utterance expresses the proposition that lemon pie is tasty to Rachel, and Lupo's utterance expresses the proposition that lemon pie is tasty to Lupo. This analysis predicts that neither speaker is at fault, as the two

[7] The replies "No it's not" and "No it's not, it's dreadful" might feel more natural and work for our purposes just as well.

[8] See Kölbel (2004) and Lasersohn (2005). For objections to this idea of faultless disagreement see Glanzberg (2007) and Stojanovic (2007).

propositions are mutually consistent, and both Rachel and Lupo appear to be in suitable positions to assert them. However, the intuition that the interlocuters are disagreeing is not accommodated by the analysis. If the contextualist is correct, then Rachel and Lupo are "talking past each other" as each is reporting something about themselves, much like in the following dialogue:

(5) *Rachel*: I live in London.
Lupo: # No, I don't live in London.[9]

In (5), Rachel and Lupo are in no disagreement, as Rachel living in London is consistent with Lupo living elsewhere. Further, Lupo's response appears infelicitous because his negation of Rachel's statement seems unwarranted and not supported by the rest of Lupo's statement. If the simple contextualist analysis of (4) is on track, then (4) should involve no disagreement, and seem problematic in much the same way as (5) does. But this is not the case – the dialogue in (4) has no apparent felicitousness problems and Rachel and Lupo appear to genuinely disagree. Therefore, the simple contextualist analysis seems to run into problems with even simple and mundane dialogues involving disagreement.

What about more sophisticated contextual theories? The simple theory treats taste predicates much like indexicals in the sense that context saturates the implicit perspective argument with the speaker's perspective. But as discussed in the previous section, in some cases the relevant perspective does not appear to be the speaker's and a more sophisticated theory that allows context to fix the perspective parameter differently is required. On these more sophisticated forms of contextualism, the context sometimes saturates this perspective argument with the perspective of some subject or set of subjects other than the speaker. Could such a theory work here?

[9] I use the hashtag to denote pragmatic infelicity.

To see if such a more sophisticated contextualist theory could work for (4), let us consider the possible perspectives that may play a role in such a dialogue. Three perspectives come to mind – Rachel's perspective, Lupo's perspective and the perspective of the group composed of Rachel and Lupo. The three perspectives give rise to three readings of the dialogue parts – let us term them respectively the R-reading, the L-reading, and the R+L-reading.

Before considering how such perspectives may figure in to an account of dialogues like (4), let me briefly discuss the notion of a perspective of a group.[10] What is the relation between the perspectives of individuals and the perspective of a group composed of them? This question is an instance of a more general, and complex question regarding the relation between individual and group attitudes. I will not attempt to answer either question comprehensively, I will merely assume, somewhat tentatively, that something is tasty from the perspective of a group if and only if it is tasty from the perspective of all its individuals.[11] Therefore, lemon pie is tasty from the perspective of the group composed of Rachel and Lupo if and only if it is tasty from both of their perspectives.

[10] I will follow much of the literature on taste predicates in taking this notion for granted. However, more ought to be said here about what group perspectives are and the conditions under which groups can be said to have them. The literature on other group attitudes like beliefs (e.g., Gilbert 1987, Lackey 2020), and group agency more generally (e.g., List and Pettit 2011) is a good starting point for doing so.

[11] To put the point more formally and more generally, let R be a predicate symbol that expresses a relation that individuals and groups may bear to an object, such that $R_i(o)$ denotes that individual i stands in relation R to object o, and $R_G(o)$ denotes that group G stands in relation R towards object o. We can thus formulate two conditions on R:

- *Distribution*: If $R_G(o)$, then for all $x \in G\ R_x(o)$
- *Aggregation*: If for all $x \in G\ R_x(o)$, then $R_G(o)$

Given this terminology, I am assuming that the relation of "being tasty from the perspective of", satisfies both *Distribution* and *Aggregation*.

This is not a trivial assumption. Other relations – like knowledge, and possibility given knowledge fail to satisfy this condition.[12] Further, taste predicates are arguably gradable adjectives – one thing may be more tasty or fun than another, and things may be tasty and fun to different degrees. The standard semantics for such adjectives involves an underlying ordering $\preccurlyeq$ among the objects to which they may apply, and a threshold degree $\bar{d}$ such that the adjective applies simpliciter to any object ranked at least as high as $\bar{d}$.[13] On this picture the predicate "tasty from Rachel's perspective" gives rise to an ordering $\preccurlyeq_R$ of objects and a threshold $\bar{d}_R$ such that everything tasty to a degree of at least $\bar{d}_R$, is tasty simpliciter. Similarly, "tasty from Lupo's perspective" gives rise to an ordering $\preccurlyeq_L$ and a threshold $\bar{d}_L$. If this story applies to aggregated perspectives as well, then "tasty from the perspective of Rachel and Lupo" will give rise to a third ordering $\preccurlyeq_{R+L}$ and threshold $\bar{d}_{R+L}$. Given this framework, questions regarding the relation between individual and group perspectives are questions of aggregating orderings and thresholds, and it is notoriously difficult to give good answers to such questions.[14]

Nevertheless, I will tentatively stick to this non-trivial assumption – while later addressing ways it may be relaxed without undermining the argument – and proceed to assess whether a sophisticated form of contextualism may yield an adequate analysis of

[12] Knowledge arguably violates *Distribution*. Assume that I know only p and you know only $p \rightarrow q$. Arguably, the knowledge of the group $\{me, you\}$ includes also q despite neither of us knowing q. Possibility given knowledge violates *Aggregation*. Assume as before that I know only p and you know only $p \rightarrow q$, then $\neg q$ is not ruled out by what I know, nor is it ruled out by what you know, but arguably, it is ruled out by what the group $\{me, you\}$ knows. This property of aggregation failure of epistemic possibility is addressed in von Fintel & Gillies' (2011) treatment of contextualism for epistemic modals that I draw on in the next section.

[13] See Kennedy and McNally (2005) for such a semantics for gradable predicates. See Barker (2013) for a treatment of taste predicates as (vague) gradable adjectives.

[14] In saying this I have in mind Arrow's (1963) impossibility theorem and the mountain of ensuing literature. Though Arrow's theorem is about preference aggregation, his conditions are plausible requirements (at least *prima fascia*) for the aggregation of tastiness orderings. For example, consider his Pareto condition – if x is tastier than y from the perspectives of all individuals of a group, then x must be tastier than y from the group perspective. Much more can be said here, but it will suffice at this point to appreciate that perspective aggregation may not be as straightforward as it might initially seem.

(4). Recall that analysing Lupo's assertion under the L-reading (i.e., relative to his perspective) fails to account for his disagreement with Rachel and his negation of her statement. Perhaps then the dialogue in (4) should be read under the R+L-reading, as being about what is tasty from the perspective of the group {Rachel, Lupo}, rather than about what is tasty to each of them individually? This would vindicate Lupo's response since from his perspective lemon pie is not tasty and thus – given the non-trivial assumption – it is not tasty to the group either. However, Lupo's response is only warranted given the R+L-reading of Rachel's assertion, for if Rachel is merely reporting that lemon pie is tasty from her perspective, then Lupo's negation is inappropriate and we're back to square one. But the R+L-reading of Rachel's assertion is unavailable because Rachel may be in no position to assert anything about the tastiness of lemon pie from Lupo's perspective, and could still felicitously and justifiably assert that lemon pie is tasty.

So, it appears that there is no reading of the dialogue in (4) under which (a) Rachel is in a position to make her assertion and (b) Lupo is in a position to make his assertion and (c) they genuinely disagree. If each assertion is read relative to the perspective of the asserter, then the disagreement dissolves as they are talking pas each other. If the whole dialogue is read relative to one of the interlocutors' perspectives, then the other interlocutor is at fault as they are in no position to be assertive about their peer's attitude toward lemon pie, much less to disagree with them about it. Finally, if the dialogue is read relative to the perspective of the group, then Rachel is in no position to make her assertion as she knows nothing about Lupo's attitude toward lemon pie, or so we may assume.

Notably, this problem persists even if we relax the non-trivial assumption in certain ways. The problem remains as long as the relationship between the group perspective and the perspectives of its members is such that knowing that something is tasty from the perspective of one of the members of the group is insufficient for aptly asserting that that thing is tasty from the perspective of the group. As long as this relationship holds, Rachel will not be in a position to say that lemon pie is tasty from the perspective of the group, knowing only that it is tasty from her perspective. This relationship would hold if, for example, the group perspective was determined by the average of the perspectives of its members, or by any other function of them that requires knowing more than one's own perspective for inferring anything about the group.[15]

A similar problem occurs in cases of agreement.[16] Consider for example the following exchange:

(6) *Rachel:* Lemon pie is tasty.
Lupo: You're right, lemon pie is tasty.[17]

Like in (4), reading the dialogue in (6) from the perspective of either of the interlocutors renders the other's assertion inapt. If (6) is read relative to Lupo's perspective, then Rachel would be making a claim about Lupo's attitude to lemon pie, something she is in no position to do. Conversely, if the dialogue is understood relative to Rachel's perspective, then Lupo's assent to her claim is out of place – it would be odd for him to confirm Rachel's statement when she is in a much better position than he is to make it. On the other hand, if the dialogue is read from the group's perspective, then like in (4), Rachel would be in no position to make her assertion, given the above relationship

[15] We should also be open to the possibility that group perspectives are sometimes indeterminate. It may for example be that the perspective of the group is only determinate given unanimity among its members. So, if all members of the group take lemon pie to be tasty (not tasty) then it is tasty (not tasty) from the perspective of the group, but if members disagree about the tastiness of lemon pie, then the group perspective is indeterminate. This would disallow inferences from negative individual attitudes to group attitudes – the fact that lemon pie is not tasty from Lupo's perspective would not entail that it is not tasty from the group's perspective.

[16] While the literature on epistemic modals deals with agreement as well as disagreement, I did not encounter discussions of agreement and the challenges it poses in relation to taste predicates.

[17] The replies "You're right, it's delicious" and "I agree, lemon pie is tasty" perhaps sound more natural and would work for our purposes here just as well.

between individual perspectives and group perspectives. Finally, if each assertion is read relative to the perspective of the asserter, then Lupo's assertion would suffer from an odd discontinuity, as the proposition that lemon pie is tasty from his perspective is quite unrelated from his assent to Rachel's claim (regardless of whether this assent is licensed). To see this, let us assume that Rachel and Lupo both live in London, and that Lupo is aware of this fact. Now consider the odd discontinuity in Lupo's utterance in the following dialogue:

(7) *Rachel:* I live in London.
Lupo: # You're right, I live in London.

Rachel is right, and Lupo – let us assume – is in a position to assent to her assertion, but this assent is starkly discontinuous with the rest of his utterance about him living in London. Under a solipsistic reading, (6) would have much of the oddity of (7), but it doesn't.

The dialogues in (4) and (6) thus appear to present a problem for contextualism. The contextualist maintains that the context determines the perspectives relative to which the interlocutors' assertions are to be evaluated, but the context appears unable to do so satisfactorily. There appears to be no good way of fixing the contextual parameters of the taste predicates in (4) and (6): doing linguistic justice with Lupo requires that both his and Rachel's assertions be analysed under the R+L-reading, but Rachel's assertion is only justified under the R-reading. The contextualist thus appears unable to deliver linguistic justice to our protagonists.

The problems generated by the dialogues in (4) (and to a lesser degree in (6)) are used to propel arguments in favour of relativism. Presumably, the relativist has an easier time in accounting for faultless disagreement, as according to her story, Rachel and Lupo are indeed disagreeing about the truth value of a unique, perspective-independent proposition – namely, that lemon pie is tasty. On the other hand, no party in this dialogue is at fault as the proposition is true relative to Rachel (when Rachel saturates the judge index) and false relative to Lupo (when he fills the judge coordinate). Thus, arguably, the relativist delivers the desired account for faultless disagreement easily.[18]

However, hope is not lost for the contextualist, as I will argue in the following section that an enhanced form of contextualism devised with other purposes in mind, could help account for all that needs accounting for in the problematic dialogues above.

3. Clouds of Contexts for Taste Predicates

Contextualist theories for epistemic modals face a problem similar to the one discussed in the previous section. On a contextualist view, epistemic modals quantify over a contextually-determined restricted set of possibilities, namely, those consistent with some contextually-determined body of knowledge. For example, a sentence employing an existential epistemic modal like "Rachel might be in the Lake District" expresses the proposition that Rachel being in the lake district is consistent with some body of knowledge determined by the context. However, in some dialogues no unique body of knowledge appears to do justice to our linguistic intuitions, and it is thus unclear how the context is to deliver the goods the contextualist expects it to deliver. Like in the case of taste predicates, this problem has been used as a basis for arguments for relativism.[19]

von Fintel and Gillies (2011) propose a solution to the apparent inability of context to determine the contextual parameter that the contextualist is committed to. While their solution is tailored to the epistemic modals' version of the problem, I will argue that it

[18] If you thought that was a bit too quick and are doubtful that the relativist can really accommodate the faultless disagreement intuition – especially the genuineness of the disagreement – then you are in good company. See Schaffer (2011), section 4.3.

[19] See Lasersohn (2005).

may be useful – with modest alteration – in the context of taste predicates as well.[20] In this section I will first explain how von Fintel and Gillies' theory solve the problem for epistemic modals, and then I will apply their theory to the case of taste predicates.

3.1. Epistemic Modals

To illustrate the problem for epistemic modals, let us consider the following example from von Fintel & Gillies (2011):[21]

> *"Alex is aiding Billy in the search for her keys:*
>
> (8) ***Alex**: You might have left them in the car.*
>
> *From here the conversation can take one of two paths. If Billy cannot rule out the possibility raised by Alex, an appropriate response might be:*
>
> (9) ***Billy**: You're right. Let me check.*
>
> *On the other hand, if Billy can rule out the prejacent, we find responses such as:*
>
> (10) ***Billy**: No, I still had them when we came into the house."*

The authors consider two possible bodies of knowledge that may saturate the contextual parameter relative to which the epistemic modals may be evaluated – Alex's knowledge, and the knowledge of the group composed of Alex and Billy. These bodies of knowledge give rise to two possible readings of the dialogues – the A-reading, and the A+B-reading. However, as the authors point out, neither reading is available. While the A-reading works for (8) and perhaps for (10),[22] it doesn't work for (9) as Billy is in no

[20] Schaffer (2011) suggests that von Fintel and Gillies' theory may be applicable to the case of taste predicates but does not pursue this application or demonstrate how it would work.

[21] See von Fintel & Gillies (2011), pp. 114-115.

[22] For the A-reading to work for (10) we must assume that Billy's negation applies not to the proposition purportedly expressed by Alex – that her knowledge leaves open the possibility of the keys being in the car – but to the prejacent, namely that the keys are in the car. Von Fintel & Gillies (2008) argue that understanding responses to epistemic modal statements as being about the prejacent is plausible, as the prejacent is what often ultimately matters in such contexts. Plausibly, what matters in the context of (8)-(10) is whether the

position – or so we may assume – to comment about whether Alex's knowledge is consistent with the keys being in the car. And while the A+B-reading works for (9) and (10) it cannot work for (8) as Alex is in no position – or so we may assume – to assert anything about Billy's evidential state.

Like in the case of taste predicates, the context appears to leave open contextual parameters that the contextualist requires it to determine. Von Fintel and Gillies point out that under-determination of contextual parameters is not a rare phenomenon, and that it is sometimes useful. To see this, consider their following example:[23]

> *"Billy meets Alex at a conference, and asks her:*
>
> (11) *Where are you from?*
>
> *That question is supposed, given a context, to partition answer-space according to how low-level in that context Billy wants his details about Alex to be. But notice that it's not really clear whether Billy wants to know where Alex is currently on sabbatical or where Alex teaches or where Alex went to graduate school or where Alex grew up. And—the point for us—Billy might not know what he wants to know. He just wants to know a bit more about Alex and will decide after she answers whether he got an answer to his question or not. He doesn't have to have the level of granularity sorted out before he asks the question. So, context (or context plus Billy's intentions) need not resolve the contextual ambiguity".*

von Fintel and Gillies then go on to suggest that when the conversational facts do not determine a specific body of knowledge for the evaluation of epistemic modals, utterances should be thought of as "taking place against a cloud of admissible contexts", one for each way of resolving the contextual indeterminacy. A Cloud of Contexts C may thus be thought of as the set of determinate contexts $\{c_1, c_2, \dots\}$ such that each c_i is a way

keys are in the car, or more generally, the whereabouts of the keys. Notice that this point constitutes a difference between epistemic modals and taste predicates, as the latter have no prejacents to which one may respond. Therefore, Lupo's negation in (4) must be understood as directed at the proposition expressed by Rachel.

[23] See von Fintel & Gillies (2011), p. 118.

of resolving the contextual indeterminacy, which in this case, involves the relevant body of knowledge for the epistemic modals. The authors then postulate several rules that govern the pragmatics of such conversations – conversations backed by a Cloud of Admissible Contexts – in ways that render the contextual indeterminacy conversationally useful. Three of their suggested rules will be of use to us here.

First, when an assertion takes place against a Cloud of Admissible Contexts, it does not express a unique proposition, rather, it *puts into play* multiple propositions – one for each way of resolving the contextual indeterminacy. When a proposition is "put into play" by a speaker it is available to the listener to pick up and reply to, perhaps under certain conditions. Importantly, the semantics is left intact as epistemic modals are assigned ordinary semantic values at determinate contexts. But when context is under-determined, an utterance puts into play multiple semantic values – the values of the utterance given the different ways of deciding the context. This may be summarized by the following principle:

> **TRAVELE**: Suppose the facts (linguistic and otherwise) up to t allow the groups G_1, G_2,… as resolutions of the contextual parameter for epistemic modals, these resolutions delimiting the cloud C of admissible contexts. Then an utterance of $might(\phi)$ with respect to C at t puts into play the set of propositions $P = \{[\![might(\phi)]\!]^c | c \in C\}$.[24]

Second, norms of assertion are not typically designed for cases of contextual-indeterminacy and thus must be supplemented to accommodate assertions against

[24] This is restatement of TRAVEL in von Fintel & Gillies (2011), p. 119 with slight notational changes. $might(\phi)$ is a placeholder for a (bare) epistemic modal statement like "it might rain tomorrow" in which ϕ stands for the prejacent. The $[\![\cdot]\!]^c$ function maps pieces of language to their denotation relative to a context c, such that epistemic modal statements are mapped to propositions given contexts. The rule assumes for simplicity that the group whose knowledge is relevant is the only contextually supplied information, and that each context fixes a unique group.

Clouds of Contexts. There are several options here, for example, one could require that one must be in a position to flat out assert all propositions one puts into play. Von Fintel and Gillies propose a weaker norm, namely, that one must be in a position to flat out assert at least one of the propositions one puts into play:

> **ASSERT$_E$**: Suppose an utterance of $might(\phi)$ by a speaker S puts in play the propositions P_1, P_2,…. Then S must have been in a position to flat out assert one of the P_i's.[25]

Third, von Fintel and Gillies suggest a pragmatic rule restricting the propositions to which a listener may reply, when more than one proposition is put into play by the speaker. They argue that the listener ought to reply in the most informative way available to her, that is, she must respond to the proposition that it would be maximally informative to respond to. In sum they suggest the following rule:

> **CONFIRM/DENY$_E$**: Suppose an utterance of $might(\phi)$ by a speaker S puts in play the propositions P_1, P_2,…. Then a hearer H can confirm (deny) the epistemic modal if the strongest P_i that H reasonably has an opinion about is such that H thinks it is true (false).[26]

This enhanced contextualist view yields an analysis of the dialogues in (8)-(10) that evades the problems faced by the ordinary contextualist analysis. On this picture, in uttering (8) Alex puts into play three propositions, one for each of the available readings of the epistemic modal – the A-reading, the B-reading and the A+B-reading. While Alex cannot flat out assert the B- or A+B-reading of the epistemic modal, she is in a position to flat out assert the A-reading and is therefore – given ASSERT$_E$ – licensed in her assertion of (8). Once these three propositions are put into play, CONFIRM/DENY$_E$ requires Billy to

[25] This is restatement of ASSERT in von Fintel and Gillies (2011), p. 120 with slight notational changes.
[26] This is restatement of CONFIRM/DENY in von Fintel and Gillies (2011), p. 121 with slight notational changes.

respond to the proposition to which it would be most informative to respond. Billy reasonably has an opinion regarding the B-reading and given further assumptions she also reasonably has an opinion regarding the A+B-reading.[27] This yields the A+B-readings for both (9) and (10) as desired.

According to this theory, the contextual under determination and the resulting ambiguity are a feature not a bug. They allow speakers like Alex to put into play more than they can assert. The B- and A+B-readings that Alex puts into play serve as a "probe or test or trial balloon" into Billy's evidence, by provoking her to respond to propositions that are not within Alex's assertory rights. When successful, dialogues employing epistemic modals allow interlocutors to arrive at strong conclusions relatively efficiently. In Alex and Billy's example above, (8) and (10) contribute to the conversational common ground the conclusion that the keys being in the car is inconsistent with the group's knowledge. (8) and (9) contribute the conclusion that the group's collective knowledge doesn't rule out the keys being in the car. The special ambiguity of the epistemic modals coupled with the pragmatic rules presented above allow reaching these conclusions quite easily.

[27] Von Fintel and Gillies assume plausibly that possibility given knowledge satisfies *Distribution* (see notes 10 and 11 above). Therefore, if Billy can rule out the prejacent then so can the group, and she thus reasonably has an opinion about the A+B-reading in the context of (10). However, since the authors also plausibly assume that possibility given knowledge does not satisfy Aggregation, the group may still be able to rule out the prejacent even if Billy cannot, and therefore vindicating the A+B-reading of (9) is trickier. The authors argue that in such a case Billy may defeasibly form an opinion regarding the A+B-reading, and thus (9) should be read on this reading as well (see their DEFEASIBLE CLOSURE rule on page 122). This wrinkle does not arise in my application of the theory to taste predicates because I assume the non-trivial assumption, namely, that taste predicates satisfy both Distribution and Aggregation. If the non-trivial assumption is relaxed, this complication may need to be addressed (see footnote 29 for one potential way of doing so).

3.2. Taste Predicates

Let us now apply von Fintel and Gillies' theory of Clouds of Contexts to the case of taste predicates and its problems discussed in section 2. As mentioned above the dialogues in (4) and (6) generate problems for contextualism about taste predicates that are similar in nature to the problems generated by the dialogues in (8)-(10) for contextualism about epistemic modals. In both cases the context under-determines contextual parameters that the contextualist requires it to determine – bodies of knowledge for epistemic modals and perspectives for taste predicates. In both cases, these contextual parameters depend on the selection of some group – the group whose knowledge or perspective is relevant for the evaluation of sentences employing epistemic modals or taste predicates.

Given the similarity of the problems for epistemic modals and taste predicates, and given the preceding discussion, the application of the Cloud of Contexts theory to the latter should be quite straightforward at this point. I will claim that Rachel's utterances in (4) and (6) are best understood as taking place against a Cloud of Contexts – one context for each way of deciding the perspective parameter. Like Alex, Rachel puts into play several propositions when she utters "lemon pie is tasty" – the R-reading, the L-reading, and the R+L-reading – thus generating similar ambiguity. The dialogue is then governed by adjusted versions of the three principles presented in the previous section, in a way that enables the speaker and listener to learn about each other in a conversationally efficient manner.

I will thus assume the following three principles analogous to the principles presented in the previous section:

TRAVELT: Suppose the facts (linguistic and otherwise) up to t allow the groups G_1, G_2,… as resolutions of the contextual parameter for taste predicates, these resolutions delimiting the cloud C of admissible contexts. Then an utterance of

$tasty(x)$ with respect to C at t puts into play the set of propositions $P = \{[\![tasty(x)]\!]^c | c \in C\}$.[28]

ASSERTT: Suppose an utterance of $tasty(x)$ by a speaker S puts in play the propositions P_1, P_2,…. Then S must have been in a position to flat out assert one of the P_i's.

CONFIRM/DENYT: Suppose an utterance of $tasty(x)$ by a speaker S puts in play the propositions P_1, P_2,…. Then a hearer H can confirm (deny) the taste statement if the strongest P_i that H reasonably has an opinion about is such that H thinks it is true (false).

Let us now return to the dialogues in (4) and (6). Due to TRAVELT, in both dialogues Rachel puts into play three propositions – the R-reading, the L-reading and R+L-reading, despite only being licensed to flat out assert the former. Her utterance is kosher though, given ASSERTT. Once the propositions are put into play, Lupo is required by CONFIRM/DENYT to respond to the proposition about which he reasonably has an opinion in a maximally informative manner. In both (4) and (6) Lupo reasonably has an opinion about the L-reading, and – given the non-trivial assumption – the R+L-reading. He thus responds to the latter, as this would be more informative, dissenting in (4) and assenting in (6).

On this picture, Lupo's responses in (4) and (6) are warranted as he is responding to the R+L-reading put into play by Rachel. This also renders his utterance about lemon pie continuous with his dissent/assent – both statements are to be read under the R+L-reading. Notice that this does not require Rachel's utterance to be understood under the

[28] The rule assumes for simplicity that the group whose perspective is relevant is the only contextually supplied information, and that each context fixes a unique group. $tasty(x)$ is a placeholder for a taste statement like "lemon pie is tasty" in which x stands for the purportedly tasty object. As in the previous section the $[\![\cdot]\!]^c$ function assigns pieces of language with their denotation relative to a context c.

R+L-reading, rather, all that is needed is that she put that reading into play and merely be in a position to assert the R-reading. Von Fintel and Gillies' theory thus yields a solution to the contextualist conundrums of (4) and (6) because it allows Lupo to respond to a proposition that Rachel merely put into play but was not in a position to flat out assert.

The theory also accommodates the intuitions of faultless disagreement in (4). Rachel is arguably not at fault because of the propositions she puts into play, the proposition she is licensed to flat out assert – namely, that lemon pie is tasty to her – is true and she is justified in believing it. Rachel does put into play other false propositions that she in not licensed to flat out assert – namely, the L- and R+L-readings – but this does not undermine her compliance with ASSERT$_T$, and this may be used to explain our intuitions regarding her faultlessness. Lupo is not at fault either because on this picture he asserts a true proposition that he is in a good position to assert (given the non-trivial assumption) – that lemon pie is not tasty from the perspective of the group.[29]

On the other hand, the Cloud of Contexts theory also accommodates the intuition that Rachel and Lupo are disagreeing in (4). While Lupo is not disagreeing with the proposition that Rachel was licensed to assert, he *is* disagreeing with two of the propositions she puts into play – the L-reading, and more prominently, the R+L-reading – and this accounts for our disagreement intuitions.

[29] If the non-trivial assumption is relaxed in a way that renders the group perspective indeterminate in cases of disagreement like (4), then Lupo would not be in a position to assert that lemon pie is not tasty from the perspective of the group. However, given such a relaxation, the strongest proposition that Lupo would reasonably believe to be true would be the L-reading, and thus, his assertion would be interpreted solipsistically (relative to his perspective). On this story, Rachel and Lupo are disagreeing about the L-reading which Rachel puts into play despite not being in a position to assert it. But Rachel is faultless as she is in a position to assert another proposition she puts into play, namely the R-reading. So, the Cloud of Contexts solution to the problem is robust to certain ways of relaxing the non-trivial assumption.

Admittedly, this is not an ordinary form of disagreement. Usually, when interlocutors disagree both sides flat out assert and fully endorse two opposing sides of some issue, and this is not the case in (4) according to the proposed theory. The disagreement there is about a proposition that Rachel merely puts into play, without flat out asserting, and without fully endorsing. There is therefore a sense in which the disagreement in (4) on the proposed theory is weaker than a typical disagreement. But this seems to be right. Rachel and Lupo's disagreement is unstable in an important sense. It is always possible for them to dissolve the disagreement and arrive at mutually consistent solipsistic readings of their assertions. In fact, in many contexts it is hard to imagine the conversation continuing for long without reaching such a dissolution. The dialogue in (12) is a natural continuation of (4):

(12) *Rachel*: Lemon pie is tasty.
Lupo: No, lemon pie is not tasty.
Rachel: Well, I guess it's just a matter of taste.[30]

In (12) Rachel diffuses the disagreement by essentially saying that the tastiness of lemon pie is a matter of perspective and it being tasty to her is consistent with it not being tasty to Lupo. This option of diffusing the disagreement is always available in dialogues like (4), and is not available in ordinary cases of disagreement, about say, whether lemon pie is healthy. In this sense disagreement about taste is unstable, and we should therefore expect the semantic and pragmatic stories underpinning such disagreement to be nonstandard. The Cloud of Contexts theory satisfies this expectation by telling a weaker

[30] The responses: "I was just saying that *I* like it" and "well, it's tasty to me" have the same effect of neutralizing the disagreement.

and nonstandard story about the disagreement in (4) – that it is about a proposition merely put into play, but not flat out asserted or fully endorsed by both parties.[31]

So, von Fintel and Gilllies' theory of Clouds of Contexts can be put to good use in our context of taste predicates and solve the contextualist's problems generated by the dialogues in (4) and (6). Beyond its adequacy in solving the above problems and its fittingness to the linguistic data, the application of the theory to this context may also be motivated in two other ways.

First, the striking similarity between epistemic modals and taste predicates should increase our confidence in moving pieces of theory from one domain to another. The similarity in the semantics of the two types of expressions is relatively uncontroversial and is appreciated by both contextualist like Schaffer (2011) and relativists like Stephenson (2007). Also, as the above discussion demonstrates, the contextualist faces problems of a very similar type in both domains. These theoretical similarities alone

[31] It is also possible for (4) to slip into a more objective argument about less perspectival qualities of lemon pie. Consider for example the following continuation of (4):

Rachel: Lemon pie is tasty.

Lupo: No, lemon pie is not tasty.

Rachel: But the balance between the sweetness and sourness is perfect.

Lupo: Maybe, but the texture is dreadfully slimy.

It is plausible to understand this dialogue to be about a more objective sense of "tasty". Perhaps "tasty from the perspective of some expert". On this reading Rachel and Lupo are arguing about whether the properties of lemon pie render it "tasty" in this sense. This type of dispute often arises about things for which the more objective sense of "tasty" (or of other taste predicates) is more salient. Imagine for example Rachel and Lupo arguing about whether some wine is tasty.

The Cloud of Contexts theory is useful for such dialogues as well, as in some cases the context leaves open for (4) both the continuation in (12) and the more objective continuation here. The theory can easily deal with such diversity by postulating that Rachel's first utterance puts into play a fourth reading of "tasty from the perspective of the expert". Lupo's response would then also be understood as putting into play both the objective reading and the R+W reading, leaving open both continuations to Rachel. This short analysis deserves more careful treatment on another occasion.

should motivate the application of a theory that works for one case to the other case. Especially if it works for the latter.

Second, and more importantly, the ambiguity generated by the Cloud of Contexts and the corresponding set of propositions put into play, could be understood to promote a conversational goal in the case of taste predicates, as well as in the case of epistemic modals. Von Fintel and Gillies argue that the ability to put into play multiple propositions, some of which one is not in a position to assert, allows interlocutors to consolidate their private information efficiently, and to gain new information collaboratively. In their example, and given the Cloud of Contexts theory, Alex is able to use one utterance to both divulge her information and conjecture about the group's information, inducing Billy to respond in a maximally informative way. This allows Alex and Billy to efficiently promote what is arguably the goal of their dialogue, namely, finding the keys. More generally, conversations involving epistemic modals – expressions about what might and must be given certain bodies of knowledge – are typically motivated by the goal of making informative progress and expanding the common ground. The Cloud of Contexts picture gains plausibility from the fact that it promotes this conversational goal efficiently.

Similarly, the Cloud of Contexts picture can be argued to efficiently promote a conversational goal in the case of taste predicates. In particular, on the proposed theory, the speaker is able to simultaneously divulge something about their own taste and raise a conjecture about the tastes of their interlocutors or of the relevant group, inducing the hearers to respond in a maximally informative manner. The purpose of Rachel's utterances in the above examples is plausibly twofold – both to share with Lupo something about what she likes and to find out more about what he likes, and hopefully to find something they both like. This is perhaps why Lupo is virtually obligated to respond by stating whether lemon pie is tasty. Responding with "interesting", or any other response that doesn't divulge his perspective on lemon pie would be very odd. If learning about interlocutors' tastes is among the goals of such dialogues, then the Cloud of Contexts theory demonstrates how taste predicates include a feature that allows promoting it efficiently. Instead of separately reporting that lemon pie is tasty from her perspective and asking whether it is tasty from Lupo's perspective and the perspective of the group, Rachel can just utter "lemon pie is tasty" and carry out all three operations at once. After Lupo's response, all three issues will have been settled and added to their "aesthetic common ground". Rachel and Lupo will have both learned something about each other and about the group composed of both of them, and hopefully, they will have found something in common. Therefore, the conversational efficiency of this purported feature of taste predicates generates independent motivation for accepting the Cloud of Contexts theory for taste predicates.

In sum, von Fintel and Gillies' Cloud of Contexts theory appears to solve multiple problems for contextualism about taste predicates. It vindicates Lupo's responses in (4) and (6) and it accommodates the intuitions of faultless disagreement in (4). Beyond solving these problems, the application of the theory to taste predicates is motivated by both the theoretical similarity between epistemic modals and taste predicates, and by the conversational efficiency it brings to dialogues involving such predicates.

4. Conclusion

According to von Fintel and Gillies, sometimes context underdetermines contextual parameters, and in some cases – like in epistemic modal statements – this is conversationally beneficial. In this chapter I argued that statements involving taste predicates are also among such cases. I have argued that adopting the theory of Clouds of Contexts to the analysis of taste predicates solves the contextualist's problems and accounts for our linguistic intuitions, in particular, for the intuition of faultless disagreement. Furthermore, the theory provides an elegant account of how contextual underdetermination, and ambiguity may be leveraged for conversational benefit. In the context of taste predicates, the conversational benefit is that of learning about one another's tastes. The built-in ambiguity of taste predicates enables speakers to both share facts about what is tasty to them and inquire about what is tasty to their listeners, and hopefully, to find something that is tasty to both.

Appendix A

A1. Proof of Theorem 2.1

The proof of theorem 2.1 makes use of a "*Theorem of the Alternative*" – a member of a class of theorems in linear programming that states that exactly one of two systems has a solution. It therefore requires the following notational conventions relating to vectors and matrices:

For any two n-dimensional vectors $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$:

(1) $x \geqq y$ iff $x_i \geq y_i$ for all i

(2) $x \geq y$ iff $x \geqq y$ but $x \neq y$

(3) $x > y$ iff $x_i > y_i$ for all i

When related to a vector, 0 is to be understood as a vector of zeros of the relevant dimensionality. For example, if $x = (x_1, \dots, x_n)$ then $x \geq 0$ is to be understood as $x \geq (0, \dots, 0)$ where the vector on the right side of the inequality is n-dimensional.

If A is an $m \times n$ matrix then A^T is the *Transpose* of A (the matrix $n \times m$ matrix that has as its rows A's columns and as its columns A's rows).

If A is an $m \times n$ matrix and x is an n-dimensional column vector, then Ax is their matrix-vector product.

The proof relies on the following "theorem of the alternative" from Mangasarian (1994):

> **Theorem of the alternative**:[1] Let A be an $m \times n$ matrix. Exactly one of the following alternatives holds:

[1] This theorem appears in Mangasarian (1994) p. 35 as an exercise. For a proof see Ozaki (2006), theorem 8. See also Gale (1989), theorem 2.10 (p. 49). The theorem appears in the following form:

(i) Either $Ax \geqq 0$, $x > 0$ has a solution x,

(ii) Or $A^T y \leq 0$, $y \geqq 0$ has a solution y.

Corollary 2.1: Let A be an $m \times n$ matrix. Exactly one of the following alternatives holds:

(i) Either $Ax \geqq 0$, $x > 0$ has a solution x,

(ii) Or $A^T y \leq 0$, $y \geq 0$ has a solution y

Proof: this theorem follows from the theorem of the alternative, and the following claim:

$A^T y \leq 0$, $y \geqq 0$ has a solution y iff $A^T y \leq 0$, $y \geq 0$ has a solution y.

The right to left direction is trivial. For the left to right direction assume $A^T y \leq 0$, $y \geqq 0$ has a solution y and assume by way of negation that $A^T y \leq 0$, $y \geq 0$ has no solution y. That is, the only solution of $A^T y \leq 0$, $y \geqq 0$ is $y = 0$. But then $A^T y = 0$ in contradiction to $A^T y \leq 0$.

Beyond this corollary, the proof relies on the four lemmas proven below.

Lemma 1: q is maximizable iff there exists a value tuple $v = (v_1, \dots, v_n)$ such that for every *pure* option $a_j \in A$, $\sum_{i=1}^{n} p_i v_i(q) \geq \sum_{i=1}^{n} p_i v_i(a_j)$.

Proof: the left to right direction of this claim is trivial – if according to some value tuple the expected value of q is at least as great as the expected value of any mixture, then it is also at least as great as the expected value of any pure option (trivial mixture).

Let A be an $m \times n$ matrix. Exactly one of the following alternatives holds:

(i) Either $Ax \leq 0$, $x \geqq 0$ has a solution x,
(ii) Or $A^T y \geqq 0$, $y > 0$ has a solution y

To get the version in the text replace (i) with (ii), x with y and A with A^T and vice versa.

For the other direction, assume that there exists a value tuple $v = (v_1, \dots, v_n)$ such that for every pure alternative $a_j \in A$, $\sum_{i=1}^n p_i v_i(q) \geq \sum_{i=1}^n p_i v_i(a_j)$ and assume by way of negation that q is not maximizable. That is, q doesn't have maximal expected value under any value tuple, and in particular not under v. So, for some $q' \in \Delta(A)$, $\sum_{i=1}^n p_i v_i(q) < \sum_{i=1}^n p_i v_i(q')$. Then $\sum_{i=1}^n p_i v_i(q) < \sum_{j=1}^m (q'_j \sum_{i=1}^n p_i v_i(a_j))$.[2] But the right side of the inequality is a weighted sum of the intertheoretic expected values of all pure options, so for some pure option a_k, $\sum_{i=1}^n p_i v_i(q) < \sum_{i=1}^n p_i v_i(a_k)$ in contradiction to the assumption q's expected value is greater than that of all pure options.

Lemma 2: the following three claims are equivalent:

(1) q is maximizable.

(2) For *every* value tuple v there is an n-tuple of positive real numbers $x_v = (x_1, \dots, x_n) > 0$ such that for all pure options $a_k \in A$, $\sum_{j=1}^m (q_j \sum_{i=1}^n x_i v_i(a_j)) \geq \sum_{i=1}^n x_i v_i(a_k)$.

(3) For *some* value tuple v there is an n-tuple of positive real numbers $x_v = (x_1, \dots, x_n) > 0$ such that for all pure options $a_k \in A$, $\sum_{j=1}^m (q_j \sum_{i=1}^n x_i v_i(a_j)) \geq \sum_{i=1}^n x_i v_i(a_k)$.

Proof: First let's prove that (1) and (2) are equivalent. Assume q is maximizable, then by lemma 1 there exists a value tuple $v = (v_1, \dots, v_n)$ such that for every *pure* option $a_k \in A$, $\sum_{i=1}^n p_i v_i(q) \geq \sum_{i=1}^n p_i v_i(a_k)$. Let $v' = (v'_1, \dots v'_n)$ be an arbitrary value tuple. Since $v_i, v'_i \in V_i$ they are positive affine transformations of each other (from TRANSFORMATION), there exists an n-tuple of positive real numbers $x \in \mathbb{R}^n_+$ and an n-tuple of real numbers $y \in \mathbb{R}^n$ such that $v_i(\cdot) = x_i v'_i(\cdot) + y_i$ for $i = 1, \dots, n$. Then we may restate the above maximizability condition in terms of v': for every *pure* option $a_k \in A$, $\sum_{i=1}^n p_i (x_i v'_i(q) + y_i) \geq \sum_{i=1}^n p_i (x_i v'_i(a_k) + y_i)$ which is equivalent to $\sum_{i=1}^n p_i x_i v'_i(q) \geq \sum_{i=1}^n p_i x_i v'_i(a_k)$ for

[2] To see this, notice that: $\sum_{i=1}^n p_i v_i(q') = p_1(q'_1 v_1(a_1) + \cdots + q'_m v_1(a_m)) + \cdots + p_n(q'_1 v_n(a_1) + \cdots + q'_m v_n(a_m)) = q'_1(p_1 v_1(a_1) + \cdots + p_n v_n(a_1)) + \cdots + q'_m(p_1 v_1(a_m) + \cdots + p_n v_n(a_m)) = \sum_{j=1}^m (q'_j \sum_{i=1}^n p_i v_i(a_j))$.

every *pure* option $a_k \in A$.[3] Then $z = (p_1 x_1, \dots, p_n x_n)$ is an n-tuple of positive real numbers such that for all pure options $a_k \in A$, $\sum_{j=1}^{m}\left(q_j \sum_{i=1}^{n} z_i v_i(a_j)\right) \geq \sum_{i=1}^{n} z_i v_i(a_k)$.

For the other direction, assume that for every value tuple v there is an n-tuple of positive real numbers $x_v = (x_1, \dots, x_n)$ $(x \in \mathbb{R}_{+}^{n})$ such that for all pure options $a_k \in A$, $\sum_{j=1}^{m}\left(q_j \sum_{i=1}^{n} x_i v_i(a_j)\right) \geq \sum_{i=1}^{n} x_i v_i(a_k)$. Let $v' = (v'_1, \dots, v'_n)$ be some value tuple, and let $x_{v'} = (x'_1, \dots, x'_n)$ be the corresponding n-tuple that satisfies the following condition:

(*) for all pure options $a_k \in A$, $\sum_{j=1}^{m}\left(q_j \sum_{i=1}^{n} x'_i v'_i(a_j)\right) \geq \sum_{i=1}^{n} x'_i v'_i(a_k)$

Consider the value tuple $v^* = (v_1^*, \dots, v_n^*)$ where $v_i^*(\cdot) = \frac{x'_i}{p_i} v'_i(\cdot)$ for $1 \leq i \leq n$. (*) can thus be rewritten as: for all pure options $a_k \in A$, $\sum_{j=1}^{m}\left(q_j \sum_{i=1}^{n} x'_i \frac{p_i}{x'_i} v_i^*(a_j)\right) \geq \sum_{i=1}^{n} x'_i \frac{p_i}{x'_i} v_i^*(a_k)$ which is equivalent to: $\sum_{j=1}^{m}\left(q_j \sum_{i=1}^{n} p_i v_i^*(a_j)\right) \geq \sum_{i=1}^{n} p_i v_i^*(a_k)$ for all pure options $a_k \in A$. So, by lemma 1, q is maximizable.

Second, let's prove that (1) and (3) are equivalent. From (1) to (3) is straightforward, as (1) entails (2) which entails (3). For the other direction, assume that for some value tuple v' there exists $x_{v'} = (x'_1, \dots, x'_n) > 0$ such that:

(*) for all pure options $a_k \in A$, $\sum_{j=1}^{m}\left(q_j \sum_{i=1}^{n} x'_i v'_i(a_j)\right) \geq \sum_{i=1}^{n} x'_i v'_i(a_k)$

Following the proof from the previous mention of (*) leads to the conclusion that q is maximizable. So, (3) entails (1).

Lemma 3: the following three claims are equivalent:

(1) q is maximizable.

(2) For *every* value tuple v, $M_v x_v \geqq 0$, $x_v > 0$ has a solution x_v.

(3) For *some* value tuple v, $M_v x_v \geqq 0$, $x_v > 0$ has a solution x_v.

[3] To see this, notice that $\sum_{i=1}^{n} p_i(x_i v'_i(q) + y_i) \geq \sum_{i=1}^{n} p_i(x_i v'_i(a_k) + y_i)$ is equivalent to $\sum_{i=1}^{n} p_i x_i v'_i(q) + \sum_{i=1}^{n} p_i y_i \geq \sum_{i=1}^{n} p_i x_i v'_i(a_k) + \sum_{i=1}^{n} p_i y_i$.

Where given a value tuple v, M_v is the following $m \times n$ matrix:

$$M = \begin{pmatrix} \sum_{i=1}^{m} q_i v_1(a_i) - v_1(a_1) & \cdots & \sum_{i=1}^{m} q_i v_n(a_i) - v_n(a_1) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{m} q_i v_1(a_i) - v_1(a_m) & \cdots & \sum_{i=1}^{m} q_i v_n(a_i) - v_n(a_m) \end{pmatrix}$$

And x_v is a n-dimnesional strictly positive column vector $x_v = (x_1, \dots, x_n)^T > 0$.

Proof: first let us prove the equivalence of (1) and (2). by lemma 2, q is maximizable iff for every value tuple v, there is an n-tuple of positive real numbers $x_v = (x_1, \dots, x_n)$ such that for all pure options $a_k \in A$, $\sum_{j=1}^{m}\left(q_j \sum_{i=1}^{n} x_i v_i(a_j)\right) \geq \sum_{i=1}^{n} x_i v_i(a_k)$. This is equivalent to the following system of m inequalities:[4]

(1) $x_1\left(\sum_{i=1}^{m} q_i v_1(a_i) - v_1(a_1)\right) + \cdots + x_n\left(\sum_{i=1}^{m} q_i v_n(a_i) - v_n(a_1)\right) \geq 0$

$\vdots$

(m) $x_1\left(\sum_{i=1}^{m} q_i v_1(a_i) - v_1(a_m)\right) + \cdots + x_n\left(\sum_{i=1}^{m} q_i v_n(a_i) - v_n(a_m)\right) \geq 0$

The existence of an n-tuple $x_v = (x_1, \dots, x_n)$ that satisfies the above system of m inequalities is equivalent to the condition that $M_v x \geqq 0$, $x_v > 0$ has a solution x.

The equivalence between (1) and (3) follows similarly from employing the existence claim in lemma 2 (claim 3).

Lemma 4: the following three claims are equivalent:

[4] To see this, consider the following sequence of equivalent statements: For all pure options $a_k \in A$:

(1) $\sum_{j=1}^{m}\left(q_j \sum_{i=1}^{n} x_i v_i(a_j)\right) \geq \sum_{i=1}^{n} x_i v_i(a_k)$

(2) $q_1 \sum_{i=1}^{n} x_i v_i(a_1) + \cdots + q_m \sum_{i=1}^{n} x_i v_i(a_m) \geq x_1 v_1(a_k) + \cdots + x_n v_n(a_k)$

(3) $q_1\left(x_1 v_1(a_1) + \cdots + x_n v_n(a_1)\right) + \cdots + q_m\left(x_1 v_1(a_m) + \cdots + x_n v_n(a_m)\right) \geq x_1 v_1(a_k) + \cdots + x_n v_n(a_k)$

(4) $x_1\left(q_1 v_1(a_1) + \cdots + q_m v_1(a_m)\right) + \cdots + x_n\left(q_1 v_n(a_1) + \cdots + q_m v_n(a_m)\right) \geq x_1 v_1(a_k) + \cdots + x_n v_n(a_k)$

(5) $x_1\left(q_1 v_1(a_1) + \cdots + q_m v_1(a_m) - v_1(a_k)\right) + \cdots + x_n\left(q_1 v_n(a_1) + \cdots + q_m v_n(a_m) - v_n(a_k)\right) \geq 0$

(1) q is dominated.

(2) For *every* value tuple v, $M_v^T y_v \leq 0$, $y_v \geq 0$ has a solution y_v.

(3) For *some* value tuple v, $M_v^T y_v \leq 0$, $y_v \geq 0$ has a solution y_v.

Where given a value tuple v, M_v^T is the *Transpose* of M_v as defined in lemma 3, and y_v is an m-dimensional column vector $y = (y_1, \dots y_m)^T \geq 0$.

Proof: First, let us prove that (1) entails (2). Since (2) entails (3) this will also suffice to prove that (1) entails (3).

Assume q is dominated by some $r = (r_1, \dots, r_m) \in \Delta(A)$. Then by definition, for all value tuples v:

(i) For $i = 1, .., n$: $v_i(q) = \sum_{j=1}^{m} q_j v_i(a_j) \leq \sum_{j=1}^{m} r_j v_i(a_j) = v_i(r)$ and

(ii) for some $k \in \{1, \dots, n\}$: $v_k(q) = \sum_{j=1}^{m} q_j v_k(a_j) < \sum_{j=1}^{m} r_j v_k(a_j) = v_k(r)$

For each value tuple v, this condition is equivalent to the following system of n inequalities:[5]

(1) $\quad r_1\left(\sum_{i=1}^{m} q_i v_1(a_i) - v_1(a_1)\right) + \cdots + r_m\left(\sum_{i=1}^{m} q_i v_1(a_i) - v_1(a_m)\right) \leq 0$

$\vdots$

(n) $\quad r_1\left(\sum_{i=1}^{m} q_i v_n(a_i) - v_n(a_1)\right) + \cdots + r_m\left(\sum_{i=1}^{m} q_i v_n(a_i) - v_n(a_m)\right) \leq 0$

Where at least one of the inequalities is strict.

[5] To see this, consider the following sequence of equivalent statements: For $j = 1, .., n$:

(1) $\sum_{i=1}^{m} q_i v_j(a_i) \leq \sum_{i=1}^{m} r_i v_j(a_i)$

(2) $\sum_{i=1}^{m} q_i v_j(a_i) - r_1 v_j(a_1) - \cdots - r_m v_j(a_m) \leq 0$

(3) $r_1\left(\sum_{i=1}^{m} q_i v_j(a_i) - v_j(a_1)\right) + \cdots + r_m\left(\sum_{i=1}^{m} q_i v_j(a_i) - v_j(a_m)\right) \leq 0$

If q is dominated, and there exists such a mixture $r = (r_1, \dots, r_m)$ that satisfies the above system of inequalities for every value tuple v, then for every v, $M_v^T y_v \leq 0$, $y_v \geq 0$ has a solution y_v. Namely, $y_v = r$ for all value tuples v.[6] So, (1) entails (2) and (3).

Now, let us prove that (3) entails (1). Assume that for *some* value tuple v, $M_v^T y_v \leq 0$, $y_v \geq 0$ has a solution y_v. This condition is equivalent to the following system of n inequalities:

$$\text{(1)} \qquad y_1\left(\textstyle\sum_{i=1}^m q_i v_1(a_i) - v_1(a_1)\right) + \cdots + y_m\left(\textstyle\sum_{i=1}^m q_i v_1(a_i) - v_1(a_m)\right) \leq 0$$

$$\vdots$$

$$\text{(n)} \qquad y_1\left(\textstyle\sum_{i=1}^m q_i v_n(a_i) - v_n(a_1)\right) + \cdots + y_m\left(\textstyle\sum_{i=1}^m q_i v_n(a_i) - v_n(a_m)\right) \leq 0$$

Where at least one of the inequalities is strict.

Let us denote $\bar{y} \coloneqq \sum_{i=1}^m y_i$. Dividing the above system of inequalities by $\bar{y}$ will yield the following equivalent system of n inequalities:[7]

$$\text{(1)} \qquad \frac{y_1}{\bar{y}}\left(\textstyle\sum_{i=1}^m q_i v_1(a_i) - v_1(a_1)\right) + \cdots + \frac{y_m}{\bar{y}}\left(\textstyle\sum_{i=1}^m q_i v_1(a_i) - v_1(a_m)\right) \leq 0$$

$$\vdots$$

$$\text{(n)} \qquad \frac{y_1}{\bar{y}}\left(\textstyle\sum_{i=1}^m q_i v_n(a_i) - v_n(a_1)\right) + \cdots + \frac{y_m}{\bar{y}}\left(\textstyle\sum_{i=1}^m q_i v_n(a_i) - v_n(a_m)\right) \leq 0$$

Where at least one of the inequalities is strict.

Since $r = \left(\frac{y_1}{\bar{y}}, \dots, \frac{y_m}{\bar{y}}\right)$ is a mixture, the above system of n inequalities is equivalent to the claim that q is dominated by r. Therefore, q is dominated. So, (3) entails (1) and (2) entails (1).

So, we have shown that (1) and (3) are equivalent and that (1) and (2) are equivalent, and so it follows that (2) and (3) are equivalent.

[6] To see this notice that $M^T \subseteq \mathbb{R}^{n \times m}$ is the following matrix:
$$M^T = \begin{pmatrix} \sum_{i=1}^m q_i v_1(a_i) - v_1(a_1) & \cdots & \sum_{i=1}^m q_i v_1(a_i) - v_1(a_m) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^m q_i v_n(a_i) - v_n(a_1) & \cdots & \sum_{i=1}^m q_i v_n(a_i) - v_n(a_m) \end{pmatrix}$$

[7] Notice that $\bar{y} > 0$ because $y \geq 0$.

We are now in a position to prove the theorem:

> **Theorem 2.1**: $q \in \Delta(A)$ is maximizable iff it is not dominated.

Proof: Assume $q \in \Delta(A)$ is maximizable and let v be some value tuple. Then by lemma 3, $M_v x_v \geqq 0$, $x_v > 0$ has a solution x_v. It then follows from corollary 2.1 that $M_v^T y_v \leq 0$, $y_v \geq 0$ has no solution y_v which entails by lemma 4 that q is not dominated.

Assume $q \in \Delta(A)$ is not maximizable and let v be some value tuple. Then by lemma 3, $M_v x_v \geqq 0$, $x_v > 0$ has no solution x_v. It then follows from corollary 2.1 that $M_v^T y_v \leq 0$, $y_v \geq 0$ has a solution y_v which entails by lemma 4 that q is dominated.■

A2. Proof of Claim 1

Claim 1: for every calibration $[\boldsymbol{v}]$, $E([\boldsymbol{v}])$ is purely convexly closed, i.e., $E([\boldsymbol{v}]) = Con\left(P\big(E([\boldsymbol{v}])\big)\right)$.

Proof: First, any set of alternatives $X \subseteq \Delta(A)$ is a subset of its pure convex closure: $X \subseteq Con\big(P(X)\big)$, since all elements of X are mixtures that give positive probability only to the pure components of X. So, we only need to prove that the pure convex closure of the set of maximal elements on $[\boldsymbol{v}]$ is also maximal on $[\boldsymbol{v}]$, or formally, that: $Con\left(P\big(E([\boldsymbol{v}])\big)\right) \subseteq E([\boldsymbol{v}])$.

First, let us prove that $P\big(E([\boldsymbol{v}])\big) \subseteq E([\boldsymbol{v}])$. Since $P\big(E([\boldsymbol{v}])\big) = \bigcup_{q \in E([\boldsymbol{v}])} P(q)$, it suffices to prove the following lemma:

> **Lemma 1**: if $q \in \Delta(A)$ is maximal on calibration $[\boldsymbol{v}]$, then all of q's pure components are maximal on $[\boldsymbol{v}]$. Formally: if $q \in E([\boldsymbol{v}])$ then $P(q) \subseteq E([\boldsymbol{v}])$.

Proof: assume $q \in \Delta(A)$ is maximal on $[\boldsymbol{v}]$, and let $v = (v_1, \ldots, v_n)$ be some value tuple in $[\boldsymbol{v}]$ (the choice of the value tuple is inconsequential, as the set of maximal elements is fixed across all tuples in a calibration). Assume by way of negation that not all of q's pure components are maximal on $[\boldsymbol{v}]$. That is, assume that for some $a_t \in A$ such that $q_t > 0$, a_t is not maximal on $[\boldsymbol{v}]$, and therefore: $v(a_t) < v(q)$.[8]

However, $v(q) = \sum_{i=1}^{n} p_i v_i(q) = \sum_{i=1}^{n} p_i \left(\sum_{j=1}^{m} q_j v_i(a_j)\right) = \sum_{j=1}^{m} q_j \left(\sum_{i=1}^{n} p_i v_i(a_j)\right) = \sum_{j=1}^{m} q_j v(a_j)$,[9] and so $v(q)$ is a weighted average of the values of the pure components of q. Therefore, if $v(a_t) < v(q)$ then there must be some pure alternative $a_s \neq a_t$ such that $v(a_s) > v(q)$, in contradiction to the assumption that q is maximal on $[\boldsymbol{v}]$. So, all of q's pure components are maximal on $[\boldsymbol{v}]$: $P\left(E([\boldsymbol{v}])\right) \subseteq E([\boldsymbol{v}])$.

If all of $E([\boldsymbol{v}])$'s pure components are maximal on $[\boldsymbol{v}]$, then so is any weighted average of some or all of them. So, all elements of the convex closure of $E([\boldsymbol{v}])$'s pure components are maximal on $[\boldsymbol{v}]$: $Con\left(P\left(E([\boldsymbol{v}])\right)\right) \subseteq E([\boldsymbol{v}])$. ■

A3. Proof of Claim 2

Claim 2: let $X \subseteq \Delta(A)$ be some nonempty set of alternatives. The pure convex closure of X is non-dominated if and only if there exists a calibration $[\boldsymbol{v}]$ such that $X \subseteq E([\boldsymbol{v}])$.

Proof: first, let us prove that if $Con\left(P(X)\right)$ is non-dominated then there exists a calibration $[\boldsymbol{v}]$ such that $X \subseteq E([\boldsymbol{v}])$. Let $q \in Con\left(P(X)\right)$ be some mixture that gives positive probability to *all* pure components of X, and notice that $X \subseteq Con\left(P(q)\right) = Con\left(P(X)\right)$ because q and X have the same pure components. By assumption, q is non-dominated

[8] I use the notation $v(q)$ (with no subscript) as shorthand for the inter-theoretic expected value of q on the value tuple $v = (v_1, \dots, v_n)$ and the credence vector $p = (p_1, \dots, p_n)$: $v(q) = \sum_{i=1}^{n} p_i v_i(q)$.

[9] Here's a version of this equation with the outer summation operators unpacked: $v(q) = p_1 \sum_{i=1}^{m} q_i v_1(a_i) + \dots + p_n \sum_{i=1}^{m} q_i v_n(a_i) = q_1 \sum_{i=1}^{n} p_i v_i(a_1) + \dots + q_m \sum_{i=1}^{n} p_i v_i(a_m)$.

and therefore by theorem 2.1 there exists a calibration $[\boldsymbol{v}]$ such that $q \in E([\boldsymbol{v}])$. This entails by claim 1 that $Con\big(P(q)\big) \subseteq E([\boldsymbol{v}])$, and therefore, $X \subseteq E([\boldsymbol{v}])$.

Second, let us prove the inverse: if the pure convex closure of X is *not* non-dominated, then there exists *no* calibration $[\boldsymbol{v}]$ such that $X \subseteq E([\boldsymbol{v}])$. Assume by way of negation that there exists a calibration $[\boldsymbol{v}]$ such that $X \subseteq E([\boldsymbol{v}])$. Then by claim 1 the pure convex closure of X is also maximal on $[\boldsymbol{v}]$, but then by theorem 2.1 the pure convex closure of X is non-dominated in contradiction to the assumption. ■

Appendix B

B1. Proof of Corollary 3.1:

The axioms:

1. **Weak order**: $\succcurlyeq^*$ is a complete and transitive over pairs of elements of $T \times T$.
2. **Invertibility**: if $ab \succcurlyeq^* cd$ then $dc \succcurlyeq^* ba$.
3. **Additivity**: if $ab \succcurlyeq^* a'b'$ and $bc \succcurlyeq^* b'c'$ then $ac \succcurlyeq^* a'c'$.
4. **Richness**: if $ab \succcurlyeq^* cd \succcurlyeq^* aa$ then there exists $e, e' \in T$ such that $ae \sim^* cd \sim^* e'b$.
5. **Minimality**: if $e \in E$ then for all $a \in T$, $ae \succcurlyeq^* aa$.

Corollary 3.1: if $\succcurlyeq^*$ satisfies the five axioms above, then there exists a weight function $w: T \to \mathbb{R}$ such that

(a) for $a, b, c, d \in T$, $ab \succcurlyeq^* cd$ if and only if $w(a) - w(b) \geq w(c) - w(d)$,

(b) if $e \in E$ then $w(e) = 0$,

(c) and for all $a \in T$, $w(a) \geq 0$.

Furthermore, w is unique up to proportional transformation, so if w' has the above properties then there exists a real number $a > 0$ such that $w' = aw$.

Proof: Minimality is consistent with the first four axioms (the axioms of Theorem 3.1), as nothing in the first four axioms determines whether pairs of the form ae – where a is an arbitrary triple and e is a null triple – are positive or negative. That is, the first four axioms don't determine whether $ae \succcurlyeq^* aa$ or $aa \succcurlyeq^* ae$ (or both). They do determine that at least one of those disjuncts is true (Weak order), and they do determine that ea has the opposite polarity as ae (Invertibility). But they do not determine which of the disjuncts is true. Therefore, by Theorem 3.1, the first four axioms entail the existence of a weight function $w: T \to \mathbb{R}$ that satisfies (a) and is unique up to positive affine transformation. Minimality entails further that for all null triples $e \in E$ and all triples $a \in T$, this weight function w satisfies $w(a) - w(e) \geq w(a) - w(a) = 0$, and therefore that $w(a) \geq w(e)$.

Let $e \in E$ be an arbitrary null triple and let $w_0 = w + b$ be a positive affine transformation of w where $b = -w(e)$. Due to minimality which entails null-equivalence the value of b is invariant to the choice of the null triple e. w_0 satisfies (a) because it is a positive affine transformation of w and by Theorem 3.1, all such transformations satisfy (a). w_0 satisfies (b) as it is defined to satisfy $w_0(e) = w(e) + b = 0$. Finally, w_0 satisfies (c) because as demonstrated above, for all $e \in N$ and $a \in T$ w, $w(a) \geq w(e)$, and therefore, $w(a) + b \geq w(e) + b$ which is equivalent to $w_0(a) \geq 0$. So, w_0 satisfies (a), (b) and (c).

Finally, to prove that w_0 is unique up to proportional transformation, let $w': T \to \mathbb{R}$ be some weight function that satisfies (a), (b) and (c). Since w' satisfies (a), Theorem 3.1 entails that it is a positive affine transformation of w_0, i.e., that there exist $a > 0$ and b such that $w' = aw_0 + b$. Let us assume by way of negation that w' is not a proportional transformation of w_0 and thus $b \neq 0$. But then for some null triple e, $w'(e) = aw_0 + b = 0 + b \neq 0$ in contradiction to the assumption that w' satisfies (b). Therefore, w' is a proportional transformation of w_0, and thus w_0 is unique up to proportional transformation. ■

References

Aboodi, R., Borer, A., & Enoch, D. (2008). Deontology, individualism, and uncertainty: A reply to Jackson and Smith. *The Journal of Philosophy*, *105*(5), 259–272.

Arrow, K. J. (1959). Rational choice functions and orderings. *Economica*, *26*(102), 121–127.

Arrow, K. J. (1963). *Social Choice and Individual Values*. Yale University Press.

Aumann, R. J. (1962). Utility theory without the completeness axiom. *Econometrica: Journal of the Econometric Society*, 445–462.

Bales, A. (2018). Indeterminate permissibility and choiceworthy options. *Philosophical Studies*, *175*(7), 1693–1702.

Barker, C. (2013). Negotiating Taste. *Inquiry*, *56*(2–3), 240–257.

Berker, S. (2007). Particular reasons. *Ethics*, *118*(1), 109–139.

Bossert, W., & Suzumura, K. (2010a). *Consistency, choice, and rationality*. Harvard University Press.

Bossert, W., & Suzumura, K. (2010b). *Consistency, choice, and rationality*. Harvard University Press.

Bradley, R. (2008). V—Comparing Evaluations. *Proceedings of the Aristotelian Society*, *108*(1_pt_1), 85–100.

Bradley, R. (2017). *Decision theory with a human face*. Cambridge University Press.

Brandt, F., Geist, C., & Harrenstein, P. (2016). A note on the McKelvey uncovered set and Pareto optimality. *Social Choice and Welfare*, *46*(1), 81–91.

Broome, J. (2004). Reasons. In R. J. Wallace, M. Smith, S. Scheffler, & S. Pettit (Eds.), *Reason and Value: Themes from the Moral Philosophy of Joseph Raz* (pp. 28–55). Oxford University Press.

Broome, J. (2012). *Climate matters: Ethics in a warming world (Norton global ethics series)*. WW Norton & Company.

Brown, C. (2011). Consequentialize this. *Ethics*, *121*(4), 749–771.

Brown, C. (2014). The composition of reasons. *Synthese*, *191*(5), 779–800.

Brown, J. (2019). Assertion and practical reasoning: Common or divergent epistemic standards? *Contemporary Epistemology: An Anthology*, 126–146.

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

Buchak, L. (2022). *Normative theories of rational choice: rivals to expected utility*.

Carr, J. R. (2020). Normative uncertainty without theories. *Australasian Journal of Philosophy*, *98*(4), 747–762.

Carr, J. R. (2022). The hard problem of intertheoretic comparisons. *Philosophical Studies*, *179*(4), 1401–1427.

Chang, R. (1997). Introduction. In R. Chang (Ed.), *Incommensurability, Incomparability, and Practical Reason*. Harvard University Press.

Cohen, G. A. (1989). On the currency of egalitarian justice. *Ethics*, *99*(4), 906–944.

Cotton-Barratt, O., MacAskill, W., & Ord, T. (2020). Statistical normalization methods in interpersonal and intertheoretic comparisons. *Journal of Philosophy*, *117*(2).

Crespo, I., & Fernández, R. (2011). Expressing taste in dialogue. *SEMDIAL 2011: Proceedings of the 15th Workshop on the Seantics and Pragmatics of Dialogue*, 84–93.

Dancy, J. (2000). *Practical reality*. Oxford University Press, USA.

Dancy, J. (2004). *Ethics without principles*. Clarendon Press.

Davidson, D. (1963). Actions, Reasons, and Causes. *Journal of Philosophy*, *60*(23), 685–700.

Dietrich, F. (2007). A generalised model of judgment aggregation. *Social Choice and Welfare*, *28*(4), 529–565.

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

Dietrich, F., & List, C. (2007). Arrow's theorem in judgment aggregation. *Social Choice and Welfare*, *29*(1), 19–33.

Dietrich, F., & List, C. (2013). A reason-based theory of rational choice. *Nous*, *47*(1), 104–134.

Dietrich, F., & List, C. (2018). From degrees of belief to binary beliefs: Lessons from judgment-aggregation theory. *The Journal of Philosophy*, *115*(5), 225–270.

Dokow, E., & Holzman, R. (2010). Aggregation of binary evaluations. *Journal of Economic Theory*, *145*(2), 495–511.

Douven, I. (2006). Assertion, knowledge, and rational credibility. *The Philosophical Review*, *115*(4), 449–485.

Dreze, J. (1990). *Essays on economic decisions under uncertainty*. CUP Archive.

Elga, A. (2010). Subjective probabilities should be sharp. *Philosophers*, *10*.

Enoch, D. (2010). Not just a truthometer: Taking oneself seriously (but not too seriously) in cases of peer disagreement. *Mind*, *119*(476), 953–997.

Enoch, D. (2015). Against public reason. *Oxford Studies in Political Philosophy*, *1*(20), 112–142.

Finlay, S. (2014). *Confusion of tongues: A theory of normative language*. Oxford University Press.

Finlay, S. (2019). A "Good" Explanation of Five Puzzles about Reasons. *Philosophical Perspectives*, *33*(1), 62–104.

Fogal, D., & Risberg, O. (2023). The weight of reasons. *Philosophical Studies*, *180*(9), 2573–2596.

Frankfurt, H. G. (2005). *On bullshit*. Princeton University Press.

Galaabaatar, T., & Karni, E. (2013). Subjective expected utility with incomplete preferences. *Econometrica*, *81*(1), 255–284.

Gale, D. (1989). *The theory of linear economic models*. University of Chicago press.

Gilbert, M. (1987). Modelling collective belief. *Synthese*, *73*, 185–204.

Glanzberg, M. (2007). Context, content, and relativism. *Philosophical Studies*, *136*, 1–29.

Goldschmidt, Z. (2024). Foundations for knowledge-based decision theories. *Australasian Journal of Philosophy*, 1–20.

Goldschmidt, Z., & Nissan-Rozen, I. (2021). The intrinsic value of risky prospects. *Synthese*, *198*(8), 7553–7575.

Gracely, E. J. (1996a). On the noncomparability of judgments made by different ethical theories. *Metaphilosophy*, *27*(3), 327–332.

Gracely, E. J. (1996b). On the noncomparability of judgments made by different ethical theories. *Metaphilosophy*, *27*(3), 327–332.

Greaves, H., & Cotton-Barratt, O. (2023). A Bargaining-Theoretic Approach to Moral Uncertainty. *Journal of Moral Philosophy*, *1*(aop), 1–43.

Gustafsson, J. E. (2022a). Second thoughts about my favourite theory. *Pacific Philosophical Quarterly*, *103*(3), 448–470.

Gustafsson, J. E. (2022b). Second thoughts about my favourite theory. *Pacific Philosophical Quarterly*, *103*(3), 448–470.

Gustafsson, J. E., & Torpman, O. (2014a). In defence of my favourite theory. *Pacific Philosophical Quarterly*, *95*(2), 159–174.

Gustafsson, J. E., & Torpman, O. (2014b). In defence of my favourite theory. *Pacific Philosophical Quarterly*, *95*(2), 159–174.

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

Harsanyi, J. C., & Selten, R. (1972). A generalized Nash solution for two-person bargaining games with incomplete information. *Management Science*, *18*(5-part-2), 80–106.

Hawthorne, J., & Stanley, J. (2008). Knowledge and action. *The Journal of Philosophy*, *105*(10), 571–590.

Hedden, B. (2016a). Does MITE make right? *Oxford Studies in Metaethics*, *11*, 102–128.

Hedden, B. (2016b). Does MITE make right? *Oxford Studies in Metaethics*, *11*, 102–128.

Hicks, A. (2018). Moral uncertainty and value comparison. *Oxford Studies in Metaethics*, *13*, 161–183.

Hindriks, F. (2007). The status of the knowledge account of assertion. *Linguistics and Philosophy*, *30*, 393–406.

Hudson, J. L. (1989). Subjectivization in ethics. *American Philosophical Quarterly*, *26*(3), 221–229.

Hyman, J. (1999). How knowledge works. *The Philosophical Quarterly*, *49*(197), 433–451.

Jackson, F., & Smith, M. (2006). Absolutist moral theories and uncertainty. *The Journal of Philosophy*, *103*(6), 267–283.

Johnson King, Z. (2020). *Accidentally doing the right thing*.

Johnson King, Z. (2022). Deliberation and Moral Motivation. *Oxford Studies in Metaethics, Volume 17*, *17*, 254.

Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge University Press.

Joyce, J. M. (2005). How probabilities reflect evidence. *Philosophical Perspectives*, *19*, 153–178.

Joyce, J. M. (2010). A defense of imprecise credences in inference and decision making. *Philosophical Perspectives*, *24*, 281–323.

Kaplan, D. (1989). Demonstratives. In J. Almog & J. Perry (Eds.), *Themes from Kaplan* (pp. 481–563). Oxford University Press.

Keeling, G. (2023). A dilemma for reasons additivity. *Economics & Philosophy*, *39*(1), 20–42.

Kennedy, C., & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 345–381.

Kölbel, M. (2004). III—Faultless Disagreement. *Proceedings of the Aristotelian Society*, *104*(1), 53–73.

Konek, J. (2019). Comparative Probabilities. In R. Pettigrew & J. Weisberg (Eds.), *The Open Handbook of Formal Epistemology*. PhilPapers.

Krantz, D. H., Luce, R. D. (Robert D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement. Volume 1, Additive and polynomial representations*. Academic Press.

Kratzer, A. (1981). The Notional Category of Modality. In H. J. Eikmeyer & H. Rieser (Eds.), *New Approaches in Word Semantics* (pp. 38–74). De Gruyter.

Kvanvig, J. (2009). Assertion, Knowledge, and Lotteries. In *Williamson on Knowledge* (pp. 140–160). Oxford University Press.

Lackey, J. (2007). Norms of assertion. *Noûs*, *41*(4), 594–626.

Lackey, J. (2020). *The Epistemology of Groups* (1st ed.). Oxford University Press.

Lasersohn, P. (2005). Context dependence, disagreement, and predicates of personal taste. *Linguistics and Philosophy*, *28*, 643–686.

Lassiter, D. (2011). *Measurement and modality: The scalar basis of modal semantics*. Ph. D. thesis, New York University.

Leitgeb, H. (2014). The stability theory of belief. *Philosophical Review*, *123*(2), 131–171.

Levi, I. (1974). On indeterminate probabilities. *Journal of Philosophy*, *71*(13).

Levi, I. (1990). *Hard choices: Decision making under unresolved conflict*. Cambridge University Press.

Lewis, M. (2021). The knowledge norm of assertion: keep it simple. *Synthese*, *199*(5), 12963–12984.

List, C. (2003). Are interpersonal comparisons of utility indeterminate? *Erkenntnis*, *58*(2), 229–260.

List, C., & Pettit, P. (2002). Aggregating sets of judgments: An impossibility result. *Economics & Philosophy*, *18*(1), 89–110.

List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

Lockhart, T. (2000a). *Moral uncertainty and its consequences*. Oxford University Press.

Lockhart, T. (2000b). *Moral uncertainty and its consequences*. Oxford University Press.

Lord, E., & Maguire, B. (2016). An opinionated guide to the weight of reasons. *Weighing Reasons*, *3*(8).

MacAskill, W. (2013). The infectiousness of nihilism. *Ethics*, *123*(3), 508–520.

MacAskill, W. (2016). Normative uncertainty as a voting problem. *Mind*, *125*(500), 967–1004.

MacAskill, W., Bykvist, K., & Ord, T. (2020). *Moral uncertainty*. Oxford University Press.

MacAskill, W., & Ord, T. (2020a). Why Maximize Expected Choice-Worthiness? *Noûs*, *54*(2), 327–353.

MacAskill, W., & Ord, T. (2020b). Why Maximize Expected Choice-Worthiness? *Noûs*, *54*(2), 327–353.

MacFarlane, J. (2007). Relativism and disagreement. *Philosophical Studies*, *132*, 17–31.

Maguire, B. (2016). The Value-Based Theory of Reasons. *Ergo, an Open Access Journal of Philosophy*, *3*, 233–262.

Mahtani, A. (2019). Vagueness and imprecise credence. *Vagueness and Rationality in Language Use and Cognition*, 7–30.

Mandelkern, M., & Dorst, K. (2022). *Assertion is weak*.

Mangasarian, O. L. (1994). *Nonlinear programming*. SIAM.

Markovits, J. (2010). Acting for the right reasons. *Philosophical Review*, *119*(2), 201–242.

Moss, S. (2015). Time-slice epistemology and action under indeterminacy. *Oxford Studies in Epistemology*, *5*(5), 172–193.

Nagel, T. (1970). *The Possibility of Altruism*. Princeton University Press.

Nair, S. (2016). How do Reasons Accrue? In E. Lord & B. Maguire (Eds.), *Weighing Reasons* (pp. 56–73). Oxford University Press.

Nair, S. (2021). "Adding Up" Reasons: Lessons for Reductive and Nonreductive Approaches. *Ethics*, *132*(1), 38–88.

Nash, J. F. (1950). The Bargaining Problem. *Econometrica*, *18*(2), 155–162.

Nehring, K., & Puppe, C. (2002). Strategy-proof social choice on single-peaked domains: possibility, impossibility and the space between. *Unpublished Manuscript, Department of Economics, University of California at Davis*.

Nissan-Rozen, I. (2012). Doing the best one can: A new justification for the use of lotteries. *Erasmus Journal for Philosophy and Economics*, *5*(1), 45–72.

Nissan-Rozen, I. (2015a). Against moral hedging. *Economics & Philosophy*, *31*(3), 349–369.

Nissan-Rozen, I. (2015b). Against moral hedging. *Economics & Philosophy*, *31*(3), 349–369.

Ozaki, Y. (2006). Theorems of the alternative and linear programming. *Meijo Ronso*, *6*, 33–40.

Pagin, P. (2011). Information and Assertoric Force. In J. Brown & H. Cappelen (Eds.), *Assertion: New Philosophical Essays* (pp. 97–135). Oxford University Press.

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

Pigozzi, G. (2006). Belief merging and the discursive dilemma: an argument-based account to paradoxes of judgment aggregation. *Synthese*, *152*, 285–298.

Prakken, H. (2005). A study of accrual of arguments, with applications to evidential reasoning. *Proceedings of the 10th International Conference on Artificial Intelligence and Law*, 85–94.

Ramsey, F. (1931). Truth and probability. In *The foundations of mathematics and other logical essays*. Humanities Press.

Raz, J. (1986). *The morality of freedom*. Clarendon Press.

Riedener, S. (2020). An axiomatic approach to axiological uncertainty. *Philosophical Studies*, *177*(2), 483–504.

Riedener, S. (2021a). *Uncertain values: An axiomatic approach to axiological uncertainty*. De Gruyter.

Riedener, S. (2021b). *Uncertain values: An axiomatic approach to axiological uncertainty*. De Gruyter.

Rinard, S. (2015). *A Decision Theory for Imprecise Probabilities*. *15*(7).

Rosenthal, C. (2021). What decision theory can't tell us about moral uncertainty. *Philosophical Studies*, *178*(10), 3085–3105.

Ross, J. (2006). Rejecting ethical deflationism. *Ethics*, *116*(4), 742–768.

Savage, L. J. (1972). *The foundations of statistics*. Courier Corporation.

Scanlon, T. M. (2014). *Being realistic about reasons*. Oxford University Press, USA.

Schaffer, J. (2011). Perspective in Taste Predicates and Epistemic Modals. In A. Egan & B. Weatherson (Eds.), *Epistemic Modality* (pp. 179–226). Oxford University Press.

Schoenfield, M. (2012). Chilling out on epistemic rationality: A defense of imprecise credences (and other imprecise doxastic attitudes). *Philosophical Studies*, *158*(2), 197–219.

Schoenfield, M. (2016). Moral vagueness is ontic vagueness. *Ethics*, *126*(2), 257–282.

Schroeder, M. (2007). *Slaves of the Passions*. OUP Oxford.

Schroeder, M. (2008a). Having reasons. *Philosophical Studies*, *139*, 57–71.

Schroeder, M. (2008b). Having reasons. *Philosophical Studies*, *139*, 57–71.

Schroeder, M. (2008c). Having reasons. *Philosophical Studies*, *139*, 57–71.

Schroeder, M. (2009). Means-end coherence, stringency, and subjective reasons. *Philosophical Studies*, *143*, 223–248.

Schroeder, M. (2021a). *Reasons first*. Oxford University Press.

Schroeder, M. (2021b). *Reasons first*. Oxford University Press.

Schroeder, M. (2021c). The fundamental reason for reasons fundamentalism. *Philosophical Studies*, *178*(10), 3107–3127.

Sen, A. (1993). Internal consistency of choice. *Econometrica: Journal of the Econometric Society*, 495–521.

Sepielli, A. (2009). What to do when you don't know what to do? In *Oxford Studies in Metaethics* (Vol. 4, pp. 5–28).

Sepielli, A. (2010). *'Along an imperfectly-lighted path': Practical rationality and normative uncertainty*. Rutgers The State University of New Jersey, School of Graduate Studies.

Sepielli, A. (2013). Moral uncertainty and the principle of equity among moral theories. *Philosophy and Phenomenological Research*, *86*(3), 580–589.

Sher, I. (2019). Comparative value and the weight of reasons. *Economics & Philosophy*, *35*(1), 103–158.

Sliwa, P. (2016). Moral worth and moral knowledge. *Philosophy and Phenomenological Research*, *93*(2), 393–418.

Stefánsson, H. O., & Bradley, R. (2015). How valuable are chances? *Philosophy of Science*, *82*(4), 602–625.

Stephenson, T. (2007). Judge dependence, epistemic modals, and predicates of personal taste. *Linguistics and Philosophy*, *30*, 487–525.

Stojanovic, I. (2007). Talking about taste: Disagreement, implicit arguments, and relative truth. *Linguistics and Philosophy*, *30*, 691–706.

Tarsney, C. (2018a). Intertheoretic value comparison: A modest proposal. *Journal of Moral Philosophy*, *15*(3), 324–344.

Tarsney, C. (2018b). Moral uncertainty for deontologists. *Ethical Theory and Moral Practice*, *21*, 505–520.

Tarsney, C. (2019). Normative uncertainty and social choice. *Mind*, *128*(512), 1285–1308.

Tarsney, C. (2021). Vive la Difference? Structural diversity as a challenge for metanormative theories. *Ethics*, *131*(2), 151–182.

Turri, J. (2011). The express knowledge account of assertion. *Australasian Journal of Philosophy*, *89*(1), 37–45.

von Fintel, K., & Gillies, A. S. (2008). CIA Leaks. *The Philosophical Review*, *117*(1), 77–98.

von Fintel, K., & Gillies, A. S. (2011). 'Might' Made Right. In A. Egan & B. Weatherson (Eds.), *Epistemic Modality* (pp. 108–130). Oxford University Press.

Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd ed.). Princeton University Press.

Weatherson, B. (2008). Decision making with imprecise probabilities. *Ms., Dept. of Philosophy, University of Michigan*.

Weatherson, B. (2014). Running risks morally. *Philosophical Studies*, *167*, 141–163.

Weatherson, B. (2019a). *Normative externalism*. Oxford University Press.

Weatherson, B. (2019b). *Normative externalism*. Oxford University Press.

Wedgwood, R. (2022). The Reasons Aggregation Theorem. In *Oxford Studies in Normative Ethics Volume 12* (pp. 127–148). Oxford University PressOxford.

Whiting, D. (2013). Stick to the facts: On the norms of assertion. *Erkenntnis*, *78*, 847–867.

Williams, J. R. G. (2014). Decision-making under indeterminacy. *Philosophers' Imprint*, *14*, 1–34.

Williamson, J. (2010). *In defence of objective Bayesianism*. OUP Oxford.

Williamson, T. (1996). Knowing and asserting. *The Philosophical Review*, *105*(4), 489–523.

Williamson, T. (2017). Acting on knowledge. *Knowledge First: Approaches in Epistemology and Mind*, 163–181.