



PhD-FSTM-2024-86
The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 18/12/2024 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN *INFORMATIQUE*

by

David STREIT

Born on 14 September 1989 in Stuttgart (Germany)

FROM METAETHICS TO MACHINE DECISIONS:
FORMAL MODELS OF
NORMATIVE REASONS AND THEIR APPLICATION IN
PHILOSOPHY AND
MACHINE ETHICS

Dissertation defence committee

Dr Leon van der Torre, dissertation supervisor
Professor, Université du Luxembourg

Dr Frank Hofmann, Chairman
Professor, Université du Luxembourg

Dr Jan Broersen
Professor, Utrecht University

Dr Hein Duijf
Assistant Professor, Utrecht University

Dr John Horty,
Professor, University of Maryland

From Metaethics to Machine Decisions

**Formal Models of Normative Reasons and Their Application in Philosophy
and Machine Ethics**

Acknowledgements

Were it not for the support of two people, this thesis would not have been completed. I would like to thank Aleks Knoks for his superb mentorship and critical engagement with the topics of my PhD and Andrea Puech for her amazing work helping me navigate the university bureaucracy.

I would also like to thank Arndt, Patrick, and Paul for keeping me sane during the worst stretches of the Covid-19 pandemic.

I would like to thank Camila for providing the illustrations found in this thesis.

I also would like to extend my gratitude to my colleagues in our group. My brother-in-arms Adam for keeping the spirits high during our four years together and for letting me use his name in all the unseemly thought experiments; Liuwen for her wise insights and being a friend; Vincent and Benoît for being brave enough to write papers with me (and, Benoît, for letting me use your name in thought experiments); Shohreh, Ali, Jérémie, and Réka for giving me a warm welcome during the pandemic. Pere for wine, Emil for helping me decipher Luxembourgish culture, Tomer and Alex for teaching me to teach.

Additionally, I would like to thank everyone who let me win at the foosball table or who spent their time drinking coffee with me instead of working, so thank you (in truly random order) Lisha, Rachele, Josephine, Xu, Giuseppe, Matti, Richard, and Marco.

I would also like to especially thank Shayan Weber for helping me understand the arcana of our doctoral school.

Lastly, I would like to thank my Jury, specifically the external jury members Jeff Horty and Hein Duijf who graciously accepted to assess this thesis.

Contents

1. Introduction	7
1.1. Formal Philosophy & Formal Ethics	8
1.2. Methodology	9
1.3. The Structure of the Thesis	10
1.4. Normative (Practical) Reasons: A Primer	11
1.4.1. Justifying and Requiring Reasons	13
1.4.2. The Kitchen Scale Model	13
1.5. Formal Models of Reasons	14
1.5.1. Justification Logic	14
1.5.2. Models Based on Expected Value	15
1.5.3. Horty's Default Logic Model	16
1.5.4. Dietrich & List's Decision Theoretic Model	17
2. Supererogation and Dual-Role Views of Reasons	21
2.1. Introduction	21
2.2. A Baseline Model of Reasons	22
2.3. Adding Justifying and Requiring Reasons	26
2.4. Supererogation Puzzles	30
2.4.1. The Classic Paradox of Supererogation	30
2.4.2. Supererogatory Holes and the All or Nothing Paradox	32
2.4.3. The Intransitivity Paradox	35
2.5. Modeling the Cases Without Justifying Reasons	39
2.6. Related Work	41
2.6.1. Non-Formal Approaches	41
2.6.2. Formal Approaches	45
2.7. Shortcomings	53
2.7.1. Aggregation	53
2.7.2. Holism	54
2.8. Conclusion	55
3. The Kitchen Scale Model Formalized	57
3.1. Introduction	57
3.2. Reasons and Holism	58
3.2.1. Revisiting Reasons	58
3.2.2. Holism and Atomism	60

3.2.3. Default Weights, Final Weights and a Constrained Holism	61
3.3. A (Formal) Holist Model of Weighing Reasons	61
3.3.1. Structure: Reasons, Considerations & Normative Graphs	62
3.3.2. Final Weights & Weighing Reasons	68
3.4. A Principle-Based Analysis of Reasons and Their Final Weight	71
3.4.1. Determining Final Weights	73
3.4.2. Weighing Reasons	84
3.5. Challenges	93
3.5.1. Challenges to Weighing	93
3.5.2. Challenges to Weighting	100
3.6. Justifying & Requiring Reasons	108
3.6.1. Justifying Reasons Against	111
3.6.2. The Puzzles Revisited	113
3.7. Related Work	118
3.7.1. Weighted Argumentation	119
3.7.2. Tucker’s Dual-Scale Model	120
3.7.3. Sher’s Decision Theoretic Model	123
3.8. Conclusion	124
4. Machine Ethics from First Principles	129
4.1. Introduction	129
4.2. Machine Ethics: State of the Art	129
4.3. Training on Cases and Guessing Weights	133
4.4. Genetic Algorithms	134
4.5. Evaluating the Framework	136
4.5.1. Ideal Scenario	136
4.5.2. Synthetic Data	137
4.6. Experiments and Empirical Results	138
4.7. Related Work	143
4.7.1. Freedman et al.	143
4.7.2. GenEth	146
4.7.3. Canavotto & Horty	149
4.8. Conclusion	151
5. Conclusion	155
5.1. Revisiting the Idea of Formal Ethics	155
5.2. Summing Up	156
A. Appendix	159
A.1. Updated Definitions for Chapter 2	159
A.2. Proofs for Chapter 2	159
A.3. Proofs for Chapter 3	167
A.4. Origin of the Material	182

1. Introduction

This thesis is an exercise in *formal ethics*. The concept at the heart of it is the notion of *normative (practical) reasons*. Normative reasons are an entirely everyday phenomenon. They are the kind of things that speak in favour of or against performing some action. The fact that it is raining is a reason to take an umbrella; the fact that it would be unfair to fail your students at random is a reason against failing them at random; and the fact that putting a spoon up your nose is painful is a reason against doing so. Reasons have become a mainstay in current (meta)ethics and some philosophers go as far as to claim that every ethical concept can ultimately be reduced to and expressed with normative reasons.¹ In this thesis, we will use formal tools to study the concept of reasons and apply the resulting formal models to issues in philosophy and machine ethics.

In order to undertake this project, we need to bridge three disciplines. Philosophy guides our models and shapes the issues we are interested in, logic and knowledge representation provides the formal models, and computer science provides us with machine learning algorithms and testing environments for our machine ethics framework.

In this introductory chapter, we will, in Section 1.1, discuss what formal ethics, and more generally, formal philosophy mean in the context of this thesis. In Section 1.2, we discuss the interdisciplinary aspect of this thesis and lay out the methods and results of each chapter. In Section 1.3, we give a more detailed overview of the content of the thesis. In Section 1.4, we discuss the concept of normative reasons in more detail. In Section 1.4.1, we discuss two of the main concepts of Chapter 2: requiring and justifying reasons. Requiring reasons, on this view, work towards making performing an action required, while justifying reasons work towards making an action permitted. In Section 1.4.2, we discuss the main concept of Chapter 3: the so-called kitchen scale model of reasons. The idea that reasons come with a weight that can be added up and compared to find out what it is one ought to do. Finally, in Section 1.5, we introduce existing formal models of reasons and related concepts.

¹See for example (Scanlon, 2014), (Raz, 1999), or (Schroeder, 2021).

1. Introduction

1.1. Formal Philosophy & Formal Ethics

Formal ethics is a part of *formal philosophy*. Formal philosophy is (just) philosophy, but done formally. Formal philosophers seek to elucidate or solve philosophical problems with formal tools. A well-known area of formal philosophy and the most used such tool is *philosophical logic*. Philosophers studying logic are trying to tease out what it means to *reason* and what tools we can use to figure this out. This is often considered essentially an exercise in metaphysics. On this view, a logician asks what logical rules the universe obeys or what *correct* reasoning, necessarily, looks like. We also find a second mature field of formal philosophy in formal epistemology. Formal epistemology, for the most part, tries to analyze the notions of knowledge and (justified) belief using formal tools and while parts of formal epistemology make use of logical tools, the largest part of it is based on (Bayesian) probability theory.² One way to understand what *formal ethics* tries to do is to see it as a sibling of formal epistemology. What we are after is to formally study not justified belief or knowledge, but justified actions and what one ought to do.

Unfortunately, the field of formal ethics is still in its infancy. There are two related, but distinct, approaches that can claim to represent formal ethics. The first approach tries to use formal tools to answer (first-order) ethical questions. Here, we find, for example, the field of population ethics, that, very roughly, concerns questions of how we can compare different (future) populations.³ A typical question that is asked here is whether it is better to have a large population of somewhat happy individuals or a smaller population of very happy individuals. The second approach tries to use formal tools to answer questions about ethical concepts. Here, we find (parts) of the field of deontic logic. Deontic logic, roughly, models different normative concepts and their interaction in a logical system. Early examples include logics about how we can derive ought statements from other ought statements.⁴ But the field of deontic logic comprises more than that. It is concerned with the study of *norms* that can be anything that has some kind of standard of correctness – moral judgments, but also legal requirements or etiquette. And so, while some parts of deontic logic qualify as formal ethics, other parts do not. A second field we find in this approach, at least on one reading, is decision theory. On this reading, decision theory is an attempt to explain what one ought to do given a set of values (or preferences), more or less likely outcomes, and limited information. Decision theory, on this reading, makes a *metaethical* claim about how these things and what one ought to do are related.

This thesis falls squarely in the second approach. In this thesis, we will discuss how a particular normative concept – reasons – relates to what you ought to (or have to) do. That means that this thesis is, in large parts, engaged in the same project that decision

²For an overview of this area see (Lin, 2024).

³For an overview see (Arrhenius et al., 2022).

⁴For example, in the modal logic *D* (or *SDL*) that has become known as the “standard” deontic logic.

theory is.⁵

There is another, more ambitious way, we can interpret what we are doing as formal ethics. If the philosophers who claim that all of ethics can ultimately be reduced to normative reasons are correct, then a proper model of how reasons work might provide a foundation for all of ethics. If we can model how reasons behave formally, then we can give a mathematical foundation for ethics.

1.2. Methodology

This thesis integrates methodologies from three different disciplines – philosophy, logic & knowledge representation, and computer science.⁶ In any interdisciplinary endeavour, it is fruitful to outline the methodology used.

This thesis comprises three chapters in which we present novel results, and each chapter's methods vary slightly. In Chapter 2, our approach is one we often see in philosophy.⁷ We start out with paradoxes surrounding the phenomenon of supererogation. We are presented with a set of (logically) inconsistent statements, each of which seems intuitive and difficult to give up. We then make use of the tools defeasible logic to apply a model of so-called dual-role views of reasons that, roughly, state that reasons can relate to actions in two distinct ways. They can require that an action be done or merely justify it. This model based on dual-role views then allows us to resolve the paradoxes. Unlike a similar approach that we often find in deontic logic, we did not build this model to resolve the paradoxes. The model we propose is independently motivated. Once the model is in place, however, we can show that we can use it to analyze and resolve the paradoxes formally. If our goal was to build a formal system, we would be done. But there is an additional goal we have in this chapter. We need to explain why the paradoxes (including the statements that we argue should be given up) seem so intuitive. Here, we use our model again to show that statements related to those that we argue should be given up *do hold* in our model and so, we claim, our model can also explain why the paradoxes seemed so intractable at first. That is, in Chapter 2, the methodology we follow is essentially the same as we would find in a non-formal philosophy text. The interdisciplinary element here stems from the fact that *the tools* we use are logical rather than the more standard conceptual tools of philosophy.

⁵Whether this approach competes with decision theory or complements it depends on one's view of reasons. While we think that it is natural to read it as a competing approach, we need not (and will not) commit to this here.

⁶We leave it to the reader to decide whether logic is a part of philosophy, mathematics, computer science or an independent field of study.

⁷We see a similar approach in deontic logic, where the goal is to find a logical system that can accommodate a 'paradox' in other systems. This is similar, but not quite the same, as what we intend to do.

1. Introduction

In Chapter 3, our methodology follows what we could call the standard approach of deontic logic and knowledge representation. We observe that a popular non-formal model in philosophy, the *kitchen scale model*, allows us to overcome some of the shortcomings of existing formal models of reasons. The first task in this chapter is then to develop a formal version of this model. We then evaluate different possible semantics for this model by introducing a set of constraints or principles on possible semantics and suggest various semantics that satisfy different sets of these principles. Once the model is in place, we show that we can apply it to philosophical questions. We show that we can – formally – answer challenges to the non-formal philosophical model and – provably – rebut some claims about the nature of the weight of reasons found in the non-formal philosophical literature. In this chapter, the main method used is one of logical modeling, and the interdisciplinary aspects come both from the foundations of the model and its application – both stem from philosophy.

In Chapter 4, we turn to applications in machine ethics. Here, our approach is most closely related to what we would find in computer science. We use the model of Chapter 3 as a foundation to develop a machine ethics framework on top of it. Here, we make use of machine learning tools to develop a system of ethical learning and then evaluate the framework by running a series of simulations to test its accuracy and performance. In this chapter, the main methods we use are from computer science and the interdisciplinary aspect comes from the model we base the framework on and the field in which we intend to apply it.

1.3. The Structure of the Thesis

The thesis is structured as follows. In Section 1.4, we lay the philosophical groundwork for what follows. We will introduce the concept of normative reasons as it is understood in the philosophical literature. In particular, we discuss so-called *dual-role views* of reasons which will be used in Chapter 2 and the *kitchen scale model* of weighing reasons that says, roughly, that reasons come with a weight and that the weights of reasons for and against actions determine if we should do them. This model of reasons, so far formally unexplored, will be the topic of Chapter 3.

In Section 1.5, we introduce existing formal models of reasons. These are the model by (Faroldi and Protopopescu, 2019, 2023) based on justification logic, models of reasons based on expected value and utility theory by (Sher, 2019) and (Nair, 2021), an influential model based on default logic by (Horty, 2012), and a model linking reasons to decision theory by (Dietrich and List, 2011).

In Chapter 2, we will develop a formal model of reasons inspired by dual-role views that allows us to formally tackle three infamous paradoxes surrounding supererogation. Here,

1.4. Normative (Practical) Reasons: A Primer

after discussing the shortcomings of a model without taking into account the insights of dual-role views, we propose an extension to this model that does take them into account. This allows us to then formally model the paradoxes of supererogation. We will show that our model can not only resolve all three paradoxes, but also explain their intuitive pull. Lastly, our model allows us to formally prove that dual-role views need not be accepted to accommodate one of the paradoxes. This is contrary to claims stated in the non-formal literature. We show that for this particular paradox, it is enough to accept a certain way that reasons and what one ought to do are related.

In Chapter 3, we develop a formal model of the kitchen scale model. Here, we use recent developments in *formal argumentation*. The model we present will be a *holist* model of reasons. This means that the context we find a reason in can influence their strength. We model this holist nature of the model using what we will call *Normative Graphs*. Our model then proceeds in two steps. First, we calculate the strength of a reason given the context it is found in. Second, we use this context-dependent strength of reasons to *weigh* them and determine which actions are permitted, ought to be done or forbidden. In doing so we propose several semantics for both steps. We analyze these via a *principle-based* analysis; that is, we propose a set of constraints or principles and determine which sets of principles each semantics violates or respects.

Once this model is in place, we apply it to philosophical challenges about the possibility of treating reasons as having a *weight* and *weighing* them. We argue that our model is a helpful tool here as it allows us to formally analyze and answer these challenges.

In Chapter 4, we apply the formal kitchen scale model to the field of machine ethics. Here, we use the model together with machine learning tools to *represent* and *learn* normative information based on a set of cases. We then predict what is permitted, forbidden, or ought to be done in future cases based on this information. We propose a testing infrastructure for this framework to test its predictive success. We then run several experiments testing the framework and discuss its relevance.

We conclude the thesis with a short summary of the main results. The proofs throughout the thesis can be found in Appendix A.

1.4. Normative (Practical) Reasons: A Primer

Many people hold strong opinions about what you should do; some will even present sophisticated theories about the matter. When philosophers are engaging in the latter activity, they are talking about first-order ethical theories – theories about what is good, bad, virtuous, the right or wrong thing to do, and so forth. Still, after some thousands of years of trying, no consensus has been reached about what it is that you should do.

1. Introduction

Nevertheless, philosophers have claimed that the disparate theories in first-order ethics share common features. In particular, it has been argued that most, if not all, first-order ethical theories can be restated in the language of *normative (practical) reasons*.⁸ Normative reasons have become a mainstay in current metaethics and epistemology. Practical normative reasons are usually glossed as *considerations that speak in favour of or against some action* by some agent. As in all other areas of philosophy, there is no universal consensus about what a reason consists of. We find in (Schroeder, 2021) the very general attempt to get close to the concept of a reason. Schroeder points to three “marks” that he believes are minimal requirements for reasons: they compete against each other, they are act-oriented, and they are the sorts of considerations that one can act on. For the purposes of this thesis, these marks are not concrete enough and a more specific (and more standard) definition is helpful here: (i) A reason is a fact (ii) that stands in a specific (dis)favouring relation to some action (iii) of an agent. Additionally, the (dis)favouring relation is usually taken to be gradable; some facts favour some action more than others. We will accept this definition for the purposes of this thesis.⁹ Let us illustrate how reasons are supposed to do this job. The simplest version of hedonic (act) utilitarianism states, roughly, that an action ought to be performed if, and only if, it is the action that leads to the most happiness. Couched in terms of reasons, we can rephrase this as follows: The fact that some action will lead to happiness is a reason to perform it. The more happiness it produces, the stronger the favouring relation. Then, you ought to perform the action supported by the strongest reasons. Hedonic utilitarianism is then the claim that these facts about the promotion of happiness are reasons, and only such facts are reasons.¹⁰ Similarly, ethical theories that emphasize *rights*, such as the right to bodily autonomy, can be expressed in terms of reasons. Here, the fact that some action would violate someone’s bodily autonomy is taken to be a (particularly strong) reason against performing the action.

The importance of the concept of reasons is taken to the extreme by the *reasons first* program, which roughly states that first, the notion of a reason (or at least the favouring relation) is not analysable any further, but all other normative concepts can be defined in terms of reasons. This is most prominently discussed by (Scanlon, 1998), (Raz, 1990), and (Parfit, 2011) and has recently been defended by (Schroeder, 2021) in a book simply

⁸Prominent examples of this position include (Raz, 1999) and (Scanlon, 2014).

⁹Still, for each of these three elements we will find someone that disagrees with it. In particular, that normative reasons must be facts – true propositions – is disputed by (Schroeder, 2021) and that reasons stand in a (dis)favouring relation to *an* action is disputed by contrastivist accounts that claim that a reason always comprises a (dis)favouring relation between two actions (see (Snedegar, 2017)), where one action is (dis)favoured over another. Additionally, some authors propose additional elements that they claim concern the core of the concept of a reason. (Kiesewetter, 2017) and (Skorupski, 2010), for example, stress the importance of *time*; A reason is then a fact speaking in favour of performing some action at some point in time *t*. Nevertheless we believe that for present purposes the definition we have given above is good enough.

¹⁰Depending on one’s metaethical commitments, one might insist that these reasons are *moral reasons*, but allow for other types of reasons like prudential reasons to exist as well. We remain agnostic as to the relationship (if any) between moral reasons and other reasons (if any).

titled “Reasons First”. All this gives us good reasons to look more closely at reasons, formally and non-formally.

1.4.1. Justifying and Requiring Reasons

Recently, several philosophers have argued that the picture as sketched above is too simple. Reasons, they claim, can relate to actions in two fundamentally distinct ways; they can *require* an action be done or *justify* its performance. (Gert, 2004, 2007, 2011, 2014), from whom we borrow the terminology, speaks of justifying and requiring reasons. Requiring reasons, so to speak, push an action towards a requirement (and therefore, the omission of the action towards impermissibility), while justifying reasons push an action towards permissibility without licensing any requirement. Gert argues that these two forces represent fundamentally distinct dimensions of normativity. Similarly, (Greenspan, 2005, 2007) speaks of “positive” and “negative” reasons, where positive reasons speak in favour of an action without licensing criticism of its alternatives, whereas negative reasons do. We can also find a similar distinction in the “enticing” and “peremptory” reasons of (Dancy, 2004), “partial” and “impartial” reasons of (Parfit, 2011). This distinction has gained more traction in the last few years and so (Muñoz and Pummer, 2022) argue that justifying reasons – “prerogatives” in Muñoz terminology – can be identified with “waivable rights to oneself” and normativity is essentially two-dimensional. In a similar vein, (Tucker, 2021, forthcominga) argues that not only are there two types of reasons, the way permissions and requirements are determined is based on two different types of *scales* that compare the different types of reasons. We will return to this scale model below and discuss it (and Tucker’s take on it) in further detail in Chapter 3. To be clear, these accounts differ in detail – we will discuss some of them in Section 2.6 – but they all share the same impetus: there are two types of reasons we need to distinguish. Most of the authors above have argued on independent grounds for the existence of these two types of reasons, but accepting this distinction has also been claimed to solve some problems for holding the position that some actions are supererogatory, that is, they go *beyond the call of duty*. We will elaborate on these in Section 2.4. So far, we are aware of only a single attempt to model this distinction formally, (Mullins, 2021), but, as we will see in Section 2.6.2, this attempt has some serious faults.

1.4.2. The Kitchen Scale Model

We have said that reasons speak in favour of or against actions. But clearly, this is not all that can be said about the matter. Some reasons speak in favour of or against an action *more strongly* than others. The fact that I am hyperglycaemic is a stronger reason against eating ice cream than the fact that it might slightly increase my risk of heart attacks some years down the line, and the fact that an innocent person is being

1. Introduction

robbed is a stronger reason to attack their robber than my inborn appetite for violence. What's more, it seems that reasons can *combine* their strength. Assume you are invited for lunch at a university canteen. The only suitable option is, as you can plainly see, not going to be particularly good. It is, however, both healthy and cheap. Individually, let's say neither the fact that the food is healthy nor that it is cheap would suffice to make it the case that you should choose it. Instead, you should just drink a coffee and hope for a better meal in the evening. When taken together, however, these two reasons provide sufficient strength to actually buy the food and eat it. Here we have a case where two weaker reasons together become stronger and defeat or outweigh a stronger reason. How can we explain this phenomenon?

A common answer is an appeal to a metaphorical *weight* that reasons inherently have and that can be combined. This idea gives us the *kitchen scale model of reasons*.¹¹ The kitchen scale model of reasons says, roughly, that what we ought to do is determined by the weight of reasons for and against some possible actions. We represent possible actions by the pans of the scale. The reasons for each action go on the corresponding scale, and we should choose the action where the reasons in favour of the action have most combined weight out of all the reasons for actions and the pan tilts down.¹² This metaphor leaves it open how exactly the weight is combined, but it hints at a possibility: The weights are simply added up like physical weights.¹³ In Chapter 3, we present a formal version of the kitchen scale model and address some challenges to it.

1.5. Formal Models of Reasons

In this section, we will briefly discuss existing formal models of reasons. Specifically, we are interested in modeling reasons using justification logic (see (Faroldi and Protopopescu, 2019, 2023)), modeling reasons using probability theory & expected value (see (Sher, 2019) and subsequently (Nair, 2021)), modeling reasons using default logic (most prominently by (Horty, 2012)). Finally, (Dietrich and List, 2011), who aim to show a structural relation between reasons and preference orders in a decision framework.

1.5.1. Justification Logic

Approaches based on justification logic have so far, perhaps due to their recent publication, had little uptake in the non-formal philosophical literature. The first such attempt to model normative reasons is (Faroldi and Protopopescu, 2019). Justification logic was

¹¹We borrow the term from (Dancy, 2004), who argues *against* this picture of reasons.

¹²Breaking the metaphor somewhat, we can imagine that reasons against add negative weight to a pan.

¹³This view is defended in most detail in (Tucker, forthcominga).

introduced by Artemov (see most notably (Artemov, 2008)) in order to discuss the provability of mathematical statements while having access to explicit proof objects. Since then, it has been used to talk about epistemic justification and, recently, normative reasons. Below, we will briefly discuss the main ideas of justification logic as they relate to normative practical reasons, but we will not give a full-blown introduction to the subject.¹⁴

The idea of using justification logic for normative purposes starts with the basic idea of supplementing formulas like $O(\phi)$ – it is obligatory that ϕ – with an explicit justification – a reason for ϕ -ing. (Faroldi and Protopopescu, 2019) enrich the language just enough to be able to do this. Their formula $r : \phi$ can be read as r is a reason for ϕ . The main operation in justification logic is *application*. Originally intended to, roughly, model the concatenation of two proofs. It can be used, Faroldi and Protopopescu claim, to model aggregating two or more reasons. (Faroldi and Protopopescu, 2023) then propose ways to generate all-things-considered oughts based on these reason formulas.

This approach has some advantages. First, it allows the explicit combination and manipulation of reasons as terms in the logical language, second, a point stressed by Faroldi and Protopopescu, is that the resulting logic is *hyperintensional*. One can model a reason being a reason for ϕ without it being a reason for ψ even though ϕ and ψ are extensionally equivalent. But there are downsides. At least as the logic is currently presented, it is *too flexible* ((Faroldi and Protopopescu, 2023)). The meat of the matter for any formal model of reasons is an account of how oughts (and perhaps requirements) are derivable from reasons. The three approaches we will discuss next, offer a clear story of why this is so. In the case of Faroldi and Protopopescu, we are left with a logic that allows us to derive (i) $O(\phi)$ and $O(\neg\phi)$ at the same time as well as allows us to derive (ii) no obligation at all. Of course, this is not the last word and one can adjust the semantics so neither (i) nor (ii) holds any more, but what is missing is a principled way to go from reasons to oughts. We do not claim that this is impossible for justification logic to achieve, but for the purposes of this thesis, we do not think this model provides a good starting point.

1.5.2. Models Based on Expected Value

The second kind of formal models of reasons make use of the language of probability and expected values to reduce the notion of a reason to, broadly speaking, decision theory. This approach, made explicit first by (Sher, 2019), holds some attraction for the consequentialist school in first-order ethics and gives a coherent story of how reasons and oughts relate (see (Nair, 2021)).

¹⁴Such an introduction discussing the import of justification logic to philosophy can be found in (Artemov and Fitting, 2021).

1. Introduction

A reason for an action in this model is, roughly, a fact such that the conditional expected value of the action given the reason is higher than the unconditional expected value of the action.¹⁵ This allows for a couple of nice things all at once: First, it allows one to calculate how strong a reason is and express it numerically as the difference between the conditional expected value of an action and its unconditional expected value. It also accounts for how the context can change the strength of a reason and most importantly, it allows us to aggregate the strengths of reasons systematically. The strength of two reasons combined is just the difference of the conditional expected value of an action given that *both* reasons hold and the unconditional expected value.

While these are undoubtedly useful and helpful distinctions, the model also has some counterintuitive features. For one, a reason in this model can be basically anything. “Speaking in favour” is treated as simply raising the expected value of an action. This is, to say the least, a non-standard account. One of the examples given by (Sher, 2019) of something that can serve as a reason is the fact that *killing is wrong*; not something philosophers would readily identify as a normative reason.

While we are optimistic that one could, in principle, use such a model as a starting point to discuss the topics of the thesis, we also believe that there are other models that are a better fit; like the next one we discuss.

1.5.3. Horty’s Default Logic Model

Horty’s logic of reasons is a simplified version of Reiter’s (1980) default logic. Reasons are represented as defeasible *default rules* written as $A \rightarrow B$, which can be read as: the fact A , if it obtains, is a reason for (action) B .¹⁶ As shorthand, we will write $Premise(A \rightarrow B)$ for the formula A and $Conclusion(A \rightarrow B)$ for the formula B and extend this shorthand to sets of reasons. Note though that the symbol $\rightarrow$ in these formulas is *not* the material implication for which we will use it starting from the next chapter.

The simplest version of Horty’s model is a triple $(\mathcal{W}, \mathcal{D}, <)$ – a structure that we borrow for our model in Chapter 2.¹⁷ $\mathcal{W}$ is a set of propositions that stand for the facts out there in the world, $\mathcal{D}$ is a set of default rules, that is reasons, and $<$ is a strict partial order on $\mathcal{D}$. Oughts are derived from select subsets of reasons together with $\mathcal{W}$ called *proper scenarios*. Proper scenarios represent the reasons that are *applicable* in a given

¹⁵How exactly the expected value is calculated is a difficult question we will not address in detail here. (Sher, 2019) asserts that this can be done on the basis of what he calls “deliberative probabilities”, which are taken to be different from both subjective and objective probabilities.

¹⁶Horty also allows for other ways to read these formulas, notably epistemically, but we are here only concerned with the practical domain and so will stick to this reading throughout.

¹⁷Horty later lifts the order to the language level; a complication which is immaterial to our purposes here and that we will ignore throughout.

situation.

Horty defines three conditions for when a set $S \subseteq \mathcal{D}$ is a proper scenario. First, the reasons have to be “triggered” in the set. This means, roughly, that the premise of the reason has to be “active”. Only facts can be reasons and so only those reasons where the premise is derivable are included in the set. Formally, this means that for any reason $\delta \in S$, $Premise(\delta)$ has to follow from $\mathcal{W} \cup Conclusion(S)$. Second, the reasons in S cannot be “conflicted”, which simply means that we do not allow for inconsistencies in the set. You cannot have “applicable reasons” pulling you towards opposing actions. Formally, this means that the conclusions of the reasons in S are consistent with $\mathcal{W}$. Third, none of the reasons in S can be defeated. Here, this simply means that any reason must not be trumped by some opposing stronger reason. Formally, on Horty’s preliminary definition, this means that if there is a reason $\delta \in S$, then there is no reason δ' that is in the set of “triggered” reasons of S , such that (i) $\delta < \delta'$ (ii) $Conclusion(\delta)$ is inconsistent with $\mathcal{W}$ and $Conclusion(\delta')$.¹⁸

Oughts, $O(\cdot)$, are then derived based on these proper scenarios. Horty offers two accounts for how this is supposed to work, the *conflict account* and the *disjunctive account*. On the conflict account, $O(\phi)$ follows from $(\mathcal{W}, \mathcal{D}, <)$ just in case, ϕ follows from $\mathcal{W}$ together with the conclusions of *some* proper scenario. On the disjunctive account, ϕ follows from $(\mathcal{W}, \mathcal{D}, <)$ just in case, ϕ follows from $\mathcal{W}$ together with the conclusions of *each* proper scenario. We think overall, the disjunctive account is the more plausible, we will use an analogue of it to derive deontic statuses of actions in Chapter 2.

1.5.4. Dietrich & List’s Decision Theoretic Model

The fourth model we discuss is due to (Dietrich and List, 2011).¹⁹ They present a model of how normative reasons relate to oughts and permissions via a detour.²⁰ Dietrich and List show that given a certain interpretation of reasons and how sets of reasons are weighed, they can generate a preference relation. This preference relation can then be used, like in decision theory, to determine which of a set of possible actions ought to be performed.

Dietrich and List’s model is built on a background consisting, first, of a set of (mutually

¹⁸This is Horty’s *preliminary account of proper scenarios*. Later in the book, he proposes a more elaborate account to deal with certain issues of defeat. For the model we propose, however, these will not be a problem as our model is in some respects simpler and in others more complicated than Horty’s and these issues never arise. Thus, we ignore the more complicated account here.

¹⁹Dietrich and List also present similar ideas in (Dietrich and List, 2017a) and (Dietrich and List, 2017b). Both of these papers, however, are not as closely related to the question of this thesis, how oughts and reasons relate, as (Dietrich and List, 2011) and so we will not discuss them here.

²⁰They present their model as a model of both normative reasons and motivating reasons – reasons for which an agent acts or can act. We only focus on the normative interpretation here.

1. Introduction

exclusive and exhaustive) actions X one can choose from in some situation and a set of properties $\mathcal{P}$, some of which might be reasons for actions in some situation. If a property P holds of some action, we will say that P is true of x .²¹

Dietrich and List then establish a connection between the weighing of sets of reasons – which here simply means that the joint weight of one set is stronger than the joint weight of another set – and a preference relation on possible actions one can perform in a given case.

The joint weight of two sets of reasons M and M' is compared by a *weighing relation* $\geq$, where $M \geq M'$ means that the joint weight of the set of reasons in M is at least as large as that of M' .²² We define $<$ and $=$ as usual in terms of $\geq$.

Dietrich and List then consider connections to a preference relation $\succeq_M$ that is generated out of a set of reasons M , where $x \succeq_M y$ means that (in a choice between x and y) x is at least as choice-worthy as y , given the set of reasons M . Again, we define $\succ_M$ and $=_M$ in terms of $\succeq_M$.

Dietrich and List show that there is a close relationship between the weighing function and the preference relation, just in case two conditions on the preference relation hold. They prove that there is a weighing function such that for each $M \in \mathcal{M}$, $x \succeq_M y$ iff $\{R \in M : R \text{ is true of } x\} \geq \{R \in M : R \text{ is true of } y\}$ for all $x, y \in X$. This holds if, and only if, (i) the preference relation is indifferent between alternatives for which the same reasons are true, that is, such that for any $x, y \in X$ and any $M \in \mathcal{M}$, if $\{R \in M : R \text{ is true of } x\} = \{R \in M : R \text{ is true of } y\}$, then $x =_M y$ and (ii) reasons that do not hold of two options play no role for the preference relation between the two. That is, for $x, y \in X$ and any $M, M' \in \mathcal{M}$ with $M' \supseteq M$, if no $R \in M' \setminus M$ is true of x or y , then $x \succeq_M y$ iff $x \succeq_{M'} y$.

This result implies in particular, that one can go from a weighing relation to a preference relation. This relation, however, is not necessarily a preference *order*, in particular, it is not guaranteed to be transitive.²³

Dietrich and List show that the above result holds for weighing *orders* and preference *orders* as well, as long as we additionally assume that the set $\mathcal{P}$ is *weakly independent*. Here, weakly independent means that every $S \subseteq \mathcal{P}$ can be instantiated precisely. That is, there is an action $x \in X$, such that for every $R \in S$, R is true of x and for every reason $R' \in \mathcal{P} \setminus S$, R' is not true of x .

²¹Dietrich and List give some properties that they want possible sets of reasons to have, specifically, they want possible sets of reasons to be closed under intersection and union. We ignore this complication and simply treat the sets of possible reasons as the powerset of $\mathcal{P}$ (which is closed under intersection and union).

²²This relation does not presuppose that the weight of reasons *adds up* like in the kitchen scale model.

²³We skip the counterexample here. A proof of this statement can be found in (Dietrich and List, 2011).

For the purposes of this thesis, what Dietrich and List show is that one can go (given two constraints) from the weighing of reasons to preference functions and from these, we can get oughts via standard definitions of decision theory.²⁴

Dietrich and List's results allow us to establish a connection between decision theory and (the weighing of) reasons. However, for the purposes of this thesis, this model is not suitable. Reasons, in Dietrich and List's model remain abstract entities, we cannot compare their strength and cannot weigh them individually. This makes them difficult to incorporate into a model that takes a set of reasons as input and returns verdicts about what we ought to do, without some kind of ad-hoc formalization.²⁵

²⁴This means, we can define a choice function C , where we define the set of choiceworthy actions, the actions that are permitted in a given situation as $C(A) = \{a \in A : a \succeq b \text{ for all } b \in A\}$. We can read this as follows: if we have a choice between the actions in A , then all the actions in $C(A)$ are permitted. If there is a single action in $C(A)$, then this is the action, we ought to do. If an action is not in $C(A)$, then we are forbidden to perform it.

²⁵Dietrich and List discuss an additive weighing function that assigns each reason some weight and then treats the joint strength of sets of reasons as their combined weight. We will discuss this idea in Chapter 3 and argue that we need further machinery to make an additive weighing function work.

2. Supererogation and Dual-Role Views of Reasons

2.1. Introduction

In this chapter, we introduce a model of reasons that allows us to distinguish between two types of normative modalities: requirements, representing, roughly, what you *must* do, and *oughts*, a weaker modality about what you ought (or should) do. We then apply this model to a series of philosophical puzzles surrounding supererogation. What makes an action supererogatory is usually explained as that it *goes beyond the call of duty*. While it would be commendable (and arguably better) to perform it, you are not required to do so. Purported examples of supererogatory actions include heroic sacrifice, such as giving one's life to save a group of strangers, but also more mundane phenomena like giving a significant amount of money to charity.¹ The model we propose shows that by applying the insights of dual-role views on reasons – views that distinguish between two types of reasons, justifying and requiring reasons – we can solve some of the puzzles.

There are two central research questions that this chapter aims to answer. First, what does a faithful model of reasons look like that can account for the two deontic modalities? Second, can this model account for the puzzles surrounding supererogation?

The layout of this chapter is as follows: in Section 2.2, we present a simple model of how reasons relate to oughts that we will use as a basis for further refinement.² We show that while it can capture simple scenarios, it struggles with some cases that involve supererogatory actions. In Section 2.3, we extend the model to be able to distinguish between justifying and requiring reasons. In Section 2.4, we present three puzzling cases involving supererogation that we sharpen into three paradoxes: The *Classic Paradox* of Supererogation, the *All-or-Nothing Paradox*, and the *Intransitivity Paradox*. We show that our model can both resolve all three paradoxes and explain why the paradoxes seemed so intractable at first. In Section 2.5, we address a worry about the model; could we have solved the Classic Paradox of Supererogation in a model that does not

¹We take no stance on whether these are examples of genuine supererogatory actions. These examples are merely to set the stage.

²We do not distinguish between requirements and oughts in this simple model. We choose to talk about oughts here as this will be close to the presentation we use.

2. Supererogation and Dual-Role Views of Reasons

distinguish between justifying and requiring reasons through different modeling choices? We propose various alternatives and argue that each of these alternatives is insufficient. In Section 2.6, we compare the model to related work. In the first half of this section, we compare it to dual-role accounts in the non-formal literature. We will be specifically interested in how oughts and requirements are derived from the two types of reasons, and in whether requiring reasons are also justifying reasons.

In the second half of the section, we compare our model to other formal work. We discuss the model of reasons by (Mullins, 2021), who like us proposes a dual-role model, the supererogation account of (McNamara, 2022a) and (McNamara, 2025), and the work on permission by (Hansen, 2014), which shares structural similarities to our model.

In Section 2.7, we will discuss some shortcomings of the approach we present. In particular, we will discuss the inability to adequately represent the aggregation of reasons and how context changes can affect the strength of reasons. In Section 2.8, we briefly discuss future work and conclude.

2.2. A Baseline Model of Reasons

Let us start our investigation with a classic scenario.

Drowning Child. You’ve promised a friend to meet her for dinner. En route to the restaurant, you come across a child in a shallow pond who is going to drown, unless you help them. If you decide to rescue the child, you won’t make it to the dinner.

Two reasons are salient here: the fact that you’ve made a promise speaks in favour of going to the dinner; the fact that the child needs help speaks in favour of saving them. It also seems clear that you ought to save the child, and, thus, that it’s permissible for you not to keep the promise: the child’s dire need provides a stronger reason than the fact that you made a promise.

Our first goal is to introduce a simple model of how reasons relate to oughts and then show that some cases involving supererogation cannot be modeled in this way.

We assume as background a propositional logic L coming with all the standard connectives and equipped with the classical consequence relation $\vdash$. This allows us to regiment the example somewhat. Letting S stand for the proposition that you save the child and P for the proposition that you keep your promise, we can formalize the implicit assumptions in the short example. We collect them in a set $\mathcal{W} = \{S \rightarrow \neg P\}$.

We express reasons using the ! (bang) operator.³ Here the reason grounded in the fact that you made a promise is expressed as $!P$ and the reason you have to save the child is expressed as $!S$. We will usually read formulas like $!X$ as: “There is a reason to X ”, but still refer to the formula $!X$ itself as a reason as well.⁴ To refer to the content of a reason or a set of reasons, we use the function $Conclusion(\cdot)$. Hence, $Conclusion(!X)$ is X and for a set of reasons $M = \{!X_1, !X_2, \dots\}$, $Conclusion(M) = \{X_1, X_2, \dots\}$.

We collect these reasons in a set $\mathcal{R}$. In our example, we have thus $\mathcal{R} = \{!P, !S\}$. Lastly, we want to express the fact that the reason to save the child is stronger than the reason to keep your promise. We do this by adding a strict partial order $<$ on $\mathcal{R}$. Here $!P < !S$.

These three elements are all we need to capture examples like the one above using the notion of a *context*.

Definition 2.2.1 (Contexts (Single Type of Reason)). *A (single-type) context Δ is a structure of the form $(\mathcal{W}, \mathcal{R}, <)$, where $\mathcal{W}$ is a consistent set of propositional formulas, $\mathcal{R}$ is a finite set of !-formulas and $<$ is a strict partial order on $\mathcal{R}$.*⁵

Next, we enrich our model to capture the fact that in the scenario above, you ought to save the child. To express this, we add the operator $Ought(\cdot)$. The simplest way we could derive this ought is by way of seeing if it is a consequence of (the conclusions of) some/all maximal consistent subset(s) of the undefeated reasons. (Horty, 2012), in his simple account of defeat, defines the undefeated reasons simply as those with whom no stronger reasons conflict.⁶ Unfortunately, this definition is too simplistic for our purposes, and we have to replace it.⁷ So, instead, we make use of a slightly more complex approach, due to (Brewka, 1994) and use the notation of (Horty, 2012, Ch. 8.2). We start by defining two notions.

³Similar notation is used in (Horty, 2003), (Nair, 2016, Appendix 2), (van Fraassen, 1973), and (Hansen, 2014).

⁴We overload the terminology here to keep it tractable. Our model suppresses the fact that grounds a reason (its *premise*). It is often this fact that is referred to as the reason. Thus, our model does not allow for two reasons to have the same conclusion. This is not a big defect, and the simple model we present here can be extended to the default logic model of (Horty, 2012), that we introduced in section 1.5.3, but only at the cost of adding unnecessary complexity. For now, treat $!X$ as both an identifier of some fact that grounds a reason to X and the statement that there is some reason to X , depending on context.

⁵The constraint that the set of reasons is finite fits well with the non-formal literature. Most of the following propositions still hold for infinite sets of reasons, but the proofs are much less compact. For the remainder of this thesis, we will always assume a finite set of reasons in each context.

⁶We have seen this in Section 1.4. In our model, this would mean that in some context $(\mathcal{W}, \mathcal{R}, <)$, $!R \in \mathcal{R}$ is undefeated iff there is no $!R' \in \mathcal{R}$ consistent with $\mathcal{W}$ such that $\mathcal{W} \cup \{Conclusion(!R')\} \vdash \neg Conclusion(!R)$ and $!R < !R'$.

⁷It would lead to counterexamples to Proposition 2.4.1. Still, even in the context of formal accounts of dual-role reasons, one can find this definition (see (Mullins, 2021)). We think this is a mistake as we will discuss in Section 2.6.2.

2. Supererogation and Dual-Role Views of Reasons

Definition 2.2.2 (Active Reasons (Single-Type)). *Given a context $\Delta = (\mathcal{W}, \mathcal{R}, <)$ and $\mathcal{D} \subseteq \mathcal{R}$, let*

$$Active_{\Delta}(\mathcal{D}) = \{r \in \mathcal{R} : \mathcal{W} \cup Conclusion(\mathcal{D}) \cup Conclusion(r) \not\vdash \perp \text{ and } r \notin \mathcal{D}\}.$$

Thus, a reason r is active relative to a set of reasons $\mathcal{D}$ in case it is consistent with (the conclusions of) $\mathcal{D}$, but not (yet) in $\mathcal{D}$. The second notion we need is that of $<$ -maximal elements:

Definition 2.2.3 (Maximal Element). *Given a set of reasons $\mathcal{D}$ and a strict partial order $<$ on $\mathcal{D}$, let $Max_{<}(\mathcal{D}) = \{r \in \mathcal{D} : \text{there is no } r' \in \mathcal{D} \text{ with } r < r'\}$.*

Here is the basic idea of the Brewka-motivated approach: given a context $(\mathcal{W}, \mathcal{R}, <)$, we look at all possible ways of extending $<$ to a total order $<'$ on $\mathcal{R}$, and then, for each of those ways, we build a set of reasons whose conclusions are consistent, starting with the empty set and iteratively selecting the $<'$ -maximal element from among the reasons that are active at a given step. Our next two definitions make this idea precise.

Definition 2.2.4 (Brewka Scenarios for Totally Ordered Contexts (Single-Type)). *Let $\Delta = (\mathcal{W}, \mathcal{R}, <)$ be a context where $<$ totally orders $\mathcal{R}$. Then $\mathcal{B}$ is the Brewka scenario of Δ just in case $\mathcal{B} = \bigcup_{i \geq 0} \mathcal{B}_i$, where the sequence $\mathcal{B}_0, \mathcal{B}_1, \mathcal{B}_2, \dots$ is defined as follows:*

$$\begin{aligned} \mathcal{B}_0 &= \emptyset, \\ \mathcal{B}_{i+1} &= \begin{cases} \mathcal{B}_i & \text{if } Active_{\Delta}(\mathcal{B}_i) = \emptyset \\ \mathcal{B}_i \cup Max_{<}(Active_{\Delta}(\mathcal{B}_i)) & \text{otherwise} \end{cases} \end{aligned}$$

As an example, consider our earlier scenario. We can express the scenario *Drowning Child*, as the context $\Delta_1^* = (\{S \rightarrow \neg P\}, \{!S, !P\}, <)$, where $<$ is the smallest relation such that $!P < !S$. This is already a totally ordered relation on $\{!S, !P\}$ and so there will be a unique Brewka scenario. In the construction of any Brewka Scenario, we always start with the empty set and so $\mathcal{B}_0 = \emptyset$. We then look at $Active_{\Delta_1^*}(\mathcal{B}_0)$, which contains exactly one reason, the strongest reason in Δ_1^* , $!S$. So $\mathcal{B}_1 = \{!S\}$. We then look at $Active_{\Delta_1^*}(\mathcal{B}_1)$ and find that it is empty, as there is only one remaining reason in $\mathcal{R}$, namely $!P$ and since $\{S \rightarrow \neg P\} \cup Conclusion(\{!S\}) \vdash \neg Conclusion(!P)$, $!P$ is not in $Active_{\Delta_1^*}(\mathcal{B}_1)$. This means that for every $i \geq 1$, $\mathcal{B}_i = \{!S\}$ and so our Brewka scenario $\mathcal{B}$ is simply $\{!S\}$ as well.

Our next definition extends the notion of a Brewka scenario to contexts that are not totally ordered.

Definition 2.2.5 (Brewka Scenarios (Single-Type)). *Let $\Delta = (\mathcal{W}, \mathcal{R}, <)$ be any context. Then $\mathcal{B}$ is a Brewka scenario based on Δ just in case $\mathcal{B}$ is the Brewka scenario of some context $(\mathcal{W}, \mathcal{R}, <')$ where $<'$ is a total order extending $<$.*

We then use the notion of a Brewka scenario to define $Ought^*(\cdot)$.⁸

Definition 2.2.6 (Ought (Single-Type)). *Given a context $\Delta = (\mathcal{W}, \mathcal{R}, <)$, the formula $Ought^*(X)$ follows from Δ , written as $\Delta \vdash Ought^*(X)$, iff $\mathcal{W} \cup Conclusion(\mathcal{B}) \vdash X$ for every Brewka scenario $\mathcal{B}$ based on Δ .⁹*

We can also readily define Permission, formally $Can^*(\cdot)$:

Definition 2.2.7 (Can (Single-Type)). *Given a context $\Delta = (\mathcal{W}, \mathcal{R}, <)$, the formula $Can^*(X)$ follows from Δ , written as $\Delta \vdash Can^*(X)$, iff $\mathcal{W} \cup Conclusion(\mathcal{B}) \vdash X$ for some Brewka scenario $\mathcal{B}$ based on Δ .*

In our case, we get $\Delta_1^* \vdash Ought^*(S)$ – we ought to save the child – just as expected.

However, there are more challenging cases. Consider the following case of supererogatory heroism.

Save One or Two. Adam and Benoît are in mortal danger. You can easily and without cost to yourself save one. You can also save both, but that would involve serious harm to you: you would lose your legs.¹⁰

An intuition about this case that is shared by philosophers and laypeople alike is that in this scenario you have to save at least one person, but you are not required to save both. This is because – perhaps – you have a right to bodily integrity that trumps the need of the second person to be rescued.¹¹ Still, rescuing both is clearly really good, better,

⁸We will enrich our notion of context in Section 2.3, and change the definition of ought to take into account both justifying and requiring reasons. Thus we mark these preliminary, single-type of reasons definitions with a *.

⁹Here, and everywhere else, we use the “disjunctive account” instead of the “conflict account” (see (Horty, 2003) for a discussion). We believe in deontic settings the disjunctive account is both more tractable and more correct. (Mullins, 2021) uses a conflict account to derive requirements and obligations. We think this leads to some problems for his approach.

¹⁰The scenario comes from (Muñoz and Pummer, 2022).

¹¹There are other explanations one could give here, still, the talk of rights comes naturally in these cases; Muñoz, who defends a (non-formal) view similar to the one proposed here likes to talk of “waivable rights against oneself” (see (Muñoz, 2020)). In case the talk of supererogation clashes with your moral intuitions or commitments, treat this chapter as a good-faith model of supererogation. It aims to make supererogation both tractable and as plausible as possible, and defend it against serious challenges.

2. Supererogation and Dual-Role Views of Reasons

in fact, than just saving one person and keeping your legs. If we were to save both, we would commend you for *going beyond the call of duty*, for performing a *supererogatory* action.

To capture all of these intuitions, we would need at least two distinct modal operators. A weaker one, capturing the notion that you still *should* save both and a stronger one expressing that you *have to* save at least one person. Before we go down this path, however, we should see whether we can model the case at all with the tools we have so far.

Let A stand for the fact that I save Adam, B stand for the fact that I save Benoît, and L stand for the fact that I lose my legs. We can then model the example as the context $\Delta_2^* = (\mathcal{W}, \mathcal{R}, <)$, where $\mathcal{W} = \{A \wedge B \rightarrow L\}$, $\mathcal{R} = \{!A, !B, !\neg L\}$, and $<$ as the smallest relation such that $!A < !\neg L$ and $!B < !\neg L$.¹²

It is easy to verify that for the context Δ_2^* , we obtain $\Delta_2^* \vdash \text{Ought}^*(\neg L \wedge (A \vee B))$ but notably not $\Delta_2^* \vdash \text{Can}^*(A \wedge B)$. This means that our model claims that it is not permitted to save both Adam and Benoît and sacrifice your legs. This is not the verdict we expect, and so it appears our baseline model cannot account for the complexities that appear to be present in this case.¹³

2.3. Adding Justifying and Requiring Reasons

In this section, we extend the baseline model to capture the distinction between justifying and requiring reasons, which provides us with a better model. The first desideratum in modeling the *Save One or Two* case is that we need to allow for two deontic modalities, a stronger and a weaker one. This distinction aligns well with linguistic data about different modals and has recently received renewed attention in moral philosophy as well.¹⁴ We say that the stronger modal operator captures *requirements*, while the weaker

¹²One immediate question one might have is whether simply stipulating what the $<$ relation looks like is a legitimate move here. We return to the question of whether one can model this scenario faithfully in Section 2.5. For now, suffice it to say that we think that if one follows the intuitions about supererogation we have presented, then the right to keep one's legs grounds a particularly strong reason.

¹³One might wonder whether we have cheated by using the particular ordering relation $<$ and if we could not have gotten the results with a different choice. We address these worries below in Section 2.5.

¹⁴For the discussion of linguistic data, see e.g., (von Fintel and Iatridou, 2008), (Portner, 2009, 79–81); for its importance for ethical theory and reasons first views in particular, see (Bedke, 2011), (Silk, 2015), and (Snedegar, 2016). See also the work of (McNamara, 1996a,b), which we briefly discuss below.

one captures *oughts*.¹⁵

We achieve this goal by adding a distinction between justifying and requiring reasons to our model. This will turn out to be a quite conservative extension, nevertheless, we need to update some definitions, starting with that of a context. To make the differences to our earlier definition transparent, we typeset changes and additions in bold.

Definition 2.3.1 (Contexts). *A context Δ is a structure **of the form** $(\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$, where $\mathcal{W}$ is a consistent set of propositional formulas, $\mathcal{R}$ **and** $\mathcal{J}$ are finite sets of $!$ -formulas **such that** $\mathcal{J} \supseteq \mathcal{R}$ and $<$ is a strict partial order **on** $\mathcal{J}$.*

We say that the reasons in $\mathcal{R}$ are the *requiring* reasons in a context, and the reasons in $\mathcal{J}$ are the *justifying* reasons. The way to interpret this addition is that every requiring reason is also a justifying reason.¹⁶ This allows us to reinterpret the *Save One or Two* case in the following way: Let $\mathcal{R} = \{!A, !B\}$, as the danger to the lives of Adam and Benoît arguably *requires* you to help them, while $\mathcal{J} = \mathcal{R} \cup \{!\neg L\}$. $!\neg L$ is not in $\mathcal{R}$, it is thus a *purely justifying reason*. Here, the thinking goes that keeping your legs is your right, but this alone does not require you to do anything, it is up to you whether you make use of that right or sacrifice your legs. Furthermore, this right seems very weighty, at least if you follow the supererogation story so far; it trumps saving an extra life. Thus, we have the further condition that $!A < !\neg L$ and $!B < !\neg L$ and finally, to be able to represent the case as described, we need $\mathcal{W}$ to be the set $\{A \wedge B \rightarrow L\}$. This will be our first dual-role context $\Delta_1 = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$. This means the context Δ_1 is identical to the single-role context Δ_2^* except for the fact that we have split the set of reasons into justifying and requiring reasons.

The definition for a Brewka scenario can be restated almost verbatim, with the only difference being that we use the new context (equipped with its partial order $<$ on $\mathcal{J}$) as a base. For completeness' sake, we have included this definition in the Appendix as well (see Appendix A.1). To define Oughts, Musts and Permissions in this new setting, however, we do need to introduce further definitions.

Definition 2.3.2 (Stable Scenarios, Restricted Contexts).

Given a context $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$, a stable scenario based on Δ is any set $\mathcal{D}$ such that $\mathcal{R} \subseteq \mathcal{D} \subseteq \mathcal{J}$. Letting $<^{\mathcal{D}}$ stand for $<$ restricted to $\mathcal{D}$, we call the context $(\mathcal{W}, \mathcal{R}, \mathcal{D}, <^{\mathcal{D}})$ the restriction of Δ to $\mathcal{D}$ and denote it by $\Delta^{\mathcal{D}}$.

¹⁵For present purposes, these are technical terms and could – as we have above – equally well be expressed using the words “have to” and “should” respectively.

¹⁶This is a widely shared, but not entirely uncontroversial assumption, see for example (Muñoz and Pummer, 2022; Pummer, 2023), who argue in favour of it, but (Tucker, forthcominga) who argues against it. In any case, we believe that the formal system we propose further validates this assumption, as we will discuss in more detail in Section 2.5.

2. Supererogation and Dual-Role Views of Reasons

This definition tells us that a stable scenario contains all requiring reasons and an arbitrary collection of (purely) justifying reasons. The idea is that a stable scenario represents what an agent might use to reach her decision. Requiring reasons will always have a pull on her, so they must be included. Justifying reasons, on the other hand, are optional, just like it is up to her whether she keeps her legs, so is it up to her to consider these reasons in her decision-making.¹⁷ In our example, there are two stable scenarios for Δ_1 , $\mathcal{R} = \{!A, !B\}$ and $\mathcal{J} = \{!A, !B, !\neg L\}$.

All that is left to do is define how to derive requirements, represented by the formula $Must(\cdot)$; oughts, represented by the formula $Ought(\cdot)$; and permissions, represented by the formula $Can(\cdot)$.

Definition 2.3.3 (Oughts). *Given a context $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$, the formula $Ought(X)$ follows from Δ , written as $\Delta \vdash Ought(X)$, just in case, $\mathcal{W} \cup Conclusion(\mathcal{B}) \vdash X$ for every Brewka scenario $\mathcal{B}$ based on $\Delta^{\mathcal{R}}$.*

Oughts are thus derived purely from requiring reasons.¹⁸ And for Δ_1 it is easy to verify that $\Delta_1 \vdash Ought(A \wedge B)$ as intended.

Whereas oughts are determined on the basis of requiring reasons alone, requirements and permissions are determined on the basis of both types of reasons. The idea underlying our definitions is simple: $Must(X)$ follows from a context when, for every stable scenario based on the context, X is a consequence of all of its Brewka scenarios; and $Can(X)$ follows when, for some stable scenario based on the context, X is a consequence of one of its Brewka scenarios.

Definition 2.3.4 (Requirements). *Given a context $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$, the formula $Must(X)$ follows from it, $\Delta \vdash Must(X)$, just in case, for every stable scenario $\mathcal{D}$ based on Δ , we have $\mathcal{W} \cup Conclusion(\mathcal{B}) \vdash X$ for every Brewka scenario $\mathcal{B}$ based on $\Delta^{\mathcal{D}}$.*

Definition 2.3.5 (Permissions). *Given a context $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$, the formula $Can(X)$ follows from it, $\Delta \vdash Can(X)$, just in case, for some stable scenario $\mathcal{D}$ based on Δ , we have $\mathcal{W} \cup Conclusion(\mathcal{B}) \vdash X$ for some Brewka scenario $\mathcal{B}$ based on $\Delta^{\mathcal{D}}$.*

To illustrate this, consider again the two stable scenarios based on Δ_1 , $\{!A, !B\}$ and $\{!A, !B, !\neg L\}$, which in turn give rise to three Brewka scenarios: $\{!A, !B\}$, $\{!A, !\neg L\}$,

¹⁷This presentation is closely related to two other ways to describe this situation: waivable rights against oneself (see (Muñoz, 2020)) and excuses. The latter treats justifying reasons as excuses that might defeat obligations that would otherwise hold. This connection, however, will not be explored further here.

¹⁸The same idea can also be found in (Pummer, 2023).

2.3. Adding Justifying and Requiring Reasons

and $\{!B, !\neg L\}$. Since $A \vee B$ follows from all of them, we have $\Delta_1 \vdash Must(A \vee B)$. And since $A \wedge B$ and $\neg L$ follow from some, we have $\Delta_1 \vdash Can(A \wedge B)$ and $\Delta_1 \vdash Can(\neg L)$. You *have to* save either Adam or Benoît; you can (and ought to) save both of them; but you can also keep your legs. Thus, the model gets all the intuitions about *Save One or Two* right.

The model also gets the intuitive results for some features we want.

Observation 2.3.1. *For any context $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$,*

- *if $\Delta \vdash Must(X)$, then $\Delta \vdash Ought(X)$; and*
- *if $\Delta \vdash Ought(X)$, then $\Delta \vdash Can(X)$.*

Observation 2.3.2. *For any context $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$, none of the following hold:*

- $\Delta \vdash Ought(\perp)$
- $\Delta \vdash Must(\perp)$
- $\Delta \vdash Can(\perp)$

Observation 2.3.3. *Let $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$ be an arbitrary context. Then*

- $\Delta \vdash Ought(X \wedge Y)$ *just in case, both $\Delta \vdash Ought(X)$ and $\Delta \vdash Ought(Y)$ and*
- $\Delta \vdash Must(X \wedge Y)$ *just in case, both $\Delta \vdash Must(X)$ and $\Delta \vdash Must(Y)$.*

The proofs for each of these statements are quite straightforward, but can nevertheless be found in Appendix A.2.

Lastly, let us briefly mention how the dual-role model of reasons conservatively extends the simple model in the following sense:

Proposition 2.3.4. *Let $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$ be a dual role context and $\Delta' = (\mathcal{W}, \mathcal{R}, <')$ be a single role context (such that they have the same $\mathcal{W}$ and $\mathcal{R}$) with $<'$ being the restriction of $<$ to $\mathcal{R}$. Then:*

- $\Delta \vdash Ought(X)$ *iff* $\Delta' \vdash Ought^*(X)$

2. Supererogation and Dual-Role Views of Reasons

and additionally, if $\mathcal{J} = \mathcal{R}$, we get

- $\Delta \vdash \text{Must}(X)$ iff $\Delta \vdash \text{Ought}(X)$ iff $\Delta' \vdash \text{Ought}^*(X)$
- $\Delta \vdash \text{Can}(X)$ iff $\Delta' \vdash \text{Can}^*(X)$ iff there is a Brewka scenario $\mathcal{B}$ based on $\Delta^{\mathcal{R}}$ or on Δ or on Δ' such that $\text{Conclusion}(\mathcal{B}) \vdash X$.

The model thus behaves as we want. Requirements and Oughts are well-behaved in that they satisfy the standard postulates for deontic modalities. We have also seen that $\text{Ought}^*(\cdot)$ in the simple model is intimately related to the modalities $\text{Ought}(\cdot)$ and $\text{Must}(\cdot)$ in the dual-role model.

2.4. Supererogation Puzzles

In this section, in addition to the *Save One or Two* case, we will introduce two more puzzling cases. These three cases in turn lead to three paradoxes surrounding the concept of supererogation, which we will discuss below. We show that while the baseline model cannot account for them, the extended model that differentiates between justifying and requiring reasons can.

Supererogation has been attacked both on first-order ethical and on more formal grounds. Particularly, it has been claimed that it leads to paradoxes.¹⁹ We understand paradoxes – for the purposes of this thesis – as sets of jointly inconsistent statements, each of which seems difficult to give up. The three paradoxes we will consider are the *Classic Paradox of Supererogation*²⁰, the *Paradox of Supererogatory Holes*/the *All or Nothing paradox* and Kamm’s *Intransitivity Paradox*. The latter two paradoxes, and to a certain extent the first as well, arise naturally out of puzzling cases, which we will introduce below.

2.4.1. The Classic Paradox of Supererogation

The first puzzle is one we have already seen exemplified in the case *Save One or Two*. It is what (Muñoz, 2021) calls the *Classic Paradox of Supererogation*: How can an action that seems best not be required?

We can make this a proper paradox as follows:

¹⁹For a discussion of these paradoxes, see (Muñoz, 2021).

²⁰We take this terminology from (Muñoz, 2021).

1. The (unique) best action one could take is saving both Adam and Benoît.²¹
(intuition)
2. If an action is the (unique) best action, then this action is required.
(intuitive principle)
3. It is required to save both Adam and Benoît.
(from 1–2)
4. But (3) is false.
(intuition)

On pain of inconsistency, we need to give up one of these statements. The question is which one and why. There is an answer to this question, that we would like to call the *standard account of supererogation*. An action is supererogatory, *because* it is better than a merely permissible – that is, a permissible, but not required – action. There is a certain threshold of goodness that an action *must* have, but everything else is not required and hence supererogatory.²² Formally, this is best exemplified by a series of papers by McNamara (see (McNamara, 1996a,b, 2011)): the *Doing Well Enough* framework.²³²⁴ The standard approach works with ranked possible worlds: the higher a world’s ranking, the (morally) better it is. Requirements are determined by a threshold: if X is true in all worlds above it, it’s required that X . Permissions are duals of requirements: if it’s not the case that $\neg X$ is required, it’s permissible that X . Oughts are determined by looking at the highest-ranked worlds: if X is true in all of them, it ought to be that X . As the highest-ranked worlds are always above the threshold, requirements imply oughts, but not vice versa. So the framework accounts for the intuitive entailment relations between requirements, oughts, and permissions.²⁵

This account explains which statement we should give up (the principle), and it offers an explanation. According to the standard account, we are permitted to perform any action above a threshold of goodness, and in our case there are several such actions possible. Our model agrees on which statement we should give up, but offers a different explanation. The principle does not take into account the differing roles between justifying and requiring reasons. An action can be the best action, but the (requiring) reasons speaking in favour of it can still be trumped by stronger justifying reasons against performing the

²¹We have not given a formal account of what makes an action “the best action”. One route one could take is that an action’s value is determined by adherence to requiring reasons. We will come back to this below and discuss it in more detail in the next Chapter in Section 3.6.

²²There might be additional criteria one wants to include here, like the cost to oneself. (see for example (Smith, forthcoming)) We are agnostic about whether there is a deep normative distinction to be made. Any such addition seems compatible with the model we propose.

²³We see McNamara’s work as a representative of the dominant approach to deontic modality in philosophy and linguistics, associated, among others, with (Lewis, 1973) and (Kratzer, 2012). The difficulties that McNamara faces are symptomatic of deep problems for this dominant approach.

²⁴There is some overlap with the notion of satisficing which we will not explore here (for example, in a somewhat related setting see (Tucker, 2017)).

²⁵That is, if you’re required to X , you ought to X , and if you ought to X , then you are permitted to X .

2. Supererogation and Dual-Role Views of Reasons

action. There are, however, more paradoxes to consider and for these, we believe, our model has the upper hand.

2.4.2. Supererogatory Holes and the All or Nothing Paradox

In addition to the Classic Paradox of Supererogation, there is a related puzzle. This is especially problematic for the standard account of supererogation. (Wessels, 2015) observes a curious phenomenon: Sometimes an act that is morally better than a permissible act is impermissible. Wessels calls these *gaps in permissibility* “supererogatory holes”.²⁶ Independently, (Horton, 2017), proposes a puzzle based on the same phenomenon. Horton’s puzzle revolves around the following *All or Nothing* case:

All or Nothing. Adam and Benoît are, again, trapped in a collapsing building, but this time you will lose your legs whether you save one or both of them.

It is intuitively plausible that a world where only one person is saved is superior to one where none are. Nevertheless, walking away seems permissible, while saving only one person does not – it involves gratuitous loss of life. Thus, we have encountered a supererogatory hole. This is a puzzling case, but we can bring the problem into sharp relief by modeling it as a genuine paradox, that is a set of jointly inconsistent claims (while staying close to the presentation in (Horton, 2017)). We use the terms “morally okay”, “should” and “morally wrong” as terms of art to express (some kind of) permission, obligation or requirement, and prohibition respectively. We will formalize this paradox below with what we consider the *proper* modalities, expressible in our model.

1. It’s morally okay to save neither Adam nor Benoît. (intuition)
2. It’s morally wrong for you to save only one of them. (intuition)
3. If an act X is morally okay and an act Y is morally wrong – and X and Y are the only two available acts – one should do X , rather than Y .²⁷
(intuitive principle)
4. You should save neither Adam nor Benoît, rather than save only one of them.

²⁶Wessels also proposes a solution for these kinds of problems in a (broadly) decision theoretic framework that relies on the notion of an action being supererogatory compared to another action. For a discussion of her approach as it relates to the model proposed here, see (Knoks and Streit, 2023).

²⁷Horton suggests that his puzzle can be phrased in terms of conditional obligations (see (Horton, 2017, fn. 2)), so this is precisely what we make use of here. We define conditional obligations in our system formally below.

(from 1–3)

5. But, clearly, (4) is false.

(intuition)

Any account of supererogation must overcome this paradox. It either has to give up one of the intuitive judgments about the case, or it has to give up the intuitive principle. In both cases, we need a story about why the statements seem so intuitive. In order to address both of these points, we first formalize this paradox in our model.

In order to talk about the conditional modalities in the non-formal statement of the paradox (we talk about two actions being the only available ones or doing X *rather than* Y) we need to add them to our model. We use an idea going back to at least (Horty, 1994):

Definition 2.4.1 (Conditional Contexts). *Given a context $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$ and a formula X consistent with $\mathcal{W}$, the result of supplementing Δ with X , written as $\Delta[X]$, is the context $(\mathcal{W} \cup \{X\}, \mathcal{R}, \mathcal{J}, <)$.*

Thus, the context $\Delta[X]$ is just like Δ , except that X is now taken to be an established fact. With the notion of conditional contexts, we can specify when conditional deontic statements follow from a context:

Definition 2.4.2 (Conditional Oughts, Requirements, and Permissions). *Let Δ be an arbitrary context. Then:*

$\Delta \sim \text{Ought}(Y|X)$ just in case $\Delta[X] \sim \text{Ought}(Y)$;

$\Delta \sim \text{Must}(Y|X)$ just in case $\Delta[X] \sim \text{Must}(Y)$;

$\Delta \sim \text{Can}(Y|X)$ just in case $\Delta[X] \sim \text{Can}(Y)$.

We read a statement like $\text{Ought}(Y|X)$ as “(some agent) ought to Y given X ”. Like other conditional deontic systems, we try to stay agnostic as to the specific meaning of this conditional reading.

We are now in a position to both present the paradox the case generates formally, and translate it into our system using both justifying and requiring reasons. Let us first give a set of jointly inconsistent statements – the All or Nothing Paradox as formalized in our system.

1. $\text{Can}(\neg A \wedge \neg B)$

(intuition)

2. Supererogation and Dual-Role Views of Reasons

2. $Must(\neg((A \wedge \neg B) \vee (\neg A \wedge B)))$ (intuition)
3. For all formulas X , Y , and Z , if $Can(X)$ and $Must(\neg Y)$, then $Must(X|X \vee Y)$ (intuitive principle)
4. $Must(\neg A \wedge \neg B | (\neg A \wedge \neg B) \vee ((A \wedge \neg B) \vee (\neg A \wedge B)))$
(from 1, 2, and instantiation of 3)
5. $Must(\neg A \wedge \neg B | \neg(A \wedge B))$ (substitution of equivalent formulas, 4)
6. But, clearly, not $Must(\neg A \wedge \neg B | \neg(A \wedge B))$ (intuition)

It's worth being explicit about two assumptions in the background of our formalization. First, we take the term of art “should” in the informal presentation to express the stronger deontic modality – a *requirement*. Second, we are assuming that if an action is morally wrong, there's a requirement forbidding taking this action. Both assumptions strike us as very plausible. What our formalization does is show that the All or Nothing case generates, indeed, a genuine formal paradox that we can express using conditional deontic modalities. As they are inconsistent, one of the claims (1)-(3) or (6) has to be given up.

Let us next translate the All or Nothing Case into our formal model. We take it that again there is a reason to save Adam and a reason to save Benoît and both are requiring reasons. This gives us the context $\Delta_2 = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$, $\mathcal{W} = \{A \vee B \rightarrow L\}$, $\mathcal{R} = \{!A, !B\}$. Additionally, we have $\mathcal{J} = \mathcal{R} \cup \{!\neg L\}$. For simplicity, we take $<$ to be the empty relation.²⁸

We can easily verify, that $\Delta_2 \vdash Can(\neg A \wedge \neg B)$ and $\Delta_2 \vdash Must(\neg([A \wedge \neg B] \vee [\neg A \wedge B]))$. Furthermore, the last statement in the paradox does not hold, as $Must(\neg A \wedge \neg B | \neg[A \wedge B])$ does not follow from Δ_2 . This means that to stay consistent, we must reject the principle expressed in (3). And luckily, this is precisely what our model predicts.

The principle

$$\text{If } \Delta \vdash Can(X) \text{ and } \Delta \vdash Must(\neg Y), \text{ then } \Delta \vdash Must(X|X \vee Y).$$

²⁸We can also set $!A < !\neg L$ and $!B < !\neg L$. What is not possible is to set $!\neg L < !A$ and $!\neg L < !B$, as this will give us that no matter what, we need to save both Adam and Benoît. This thus commits us to treating the (purely) justifying reason one has to keep one's legs as at least as strong as those to save Adam and Benoît. We think this is how it should be. Even though, it is clearly overall better to sacrifice one's legs to save a person, if we take the idea presented in these cases of supererogation seriously, then the right (or justifying reason) one has to keep one's legs trumps the need of the person to be saved and thus needs to be at least as strong as the reason that need gives rise to.

is demonstrably false. This is witnessed by the context $\Delta_3 = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$ where $\mathcal{W} = \{C \rightarrow \neg D, D \rightarrow \neg E, E \rightarrow \neg C\}$, $\mathcal{R} = \{!C, !D\}$, $\mathcal{J} = \{!C, !D, !E\}$, and $!D < !C$. It's straightforward to verify that we have both $\Delta_3 \sim Can(E)$ and $\Delta_3 \sim Must(\neg D)$, while we don't have $\Delta_3 \sim Must(E|D \vee E)$.

But how can we explain the intuitive pull of the principle? We believe that it appears so natural because two very similar principles *do hold*.

Observation 2.4.1. *For any context Δ ,*

- *if $\Delta \sim Can(X)$ and $\Delta \sim Must(\neg Y)$, then $\Delta \sim Can(X|X \vee Y)$;*
- *if $\Delta \sim Ought(X)$ and $\Delta \sim Must(\neg Y)$, then $\Delta \sim Ought(X|X \vee Y)$.*

And while the validity of these two principles can explain the intuitive pull of the original one, our approach also makes clear where the original principle goes wrong: it attempts to bridge unconditional and conditional deontic statements without keeping track of the types of reasons that these statements depend on. The proofs can be found in Appendix A.2.

2.4.3. The Intransitivity Paradox

One further (in)famous puzzle about supererogation needs to be tackled by any account aiming to model it, Kamm's (1985) Intransitivity Puzzle. Again, the puzzle can be motivated using a case.

Life or Promise. You've promised a friend to meet her for dinner. En route to the restaurant, you come across a collapsing building, with Adam trapped inside. Saving him will cost you your legs.²⁹

Kamm then argues that a binary relation between actions – she calls it the “may take precedence over” relation – is not transitive. What does that mean? We think this relation is best made sense of as a conditional modality,³⁰ We believe we can faithfully translate this taking precedence over relation as the conditional modality $Can(\cdot|\cdot)$.

Letting A stand for the fact that you save Adam, P stand for the fact that you keep your promise (and go to the restaurant), and W stand for the fact that you walk away

²⁹In Kamm's original case, you lose a kidney to save a person, rather than your legs.

³⁰This seems also what Kamm herself intends in (Kamm, 1985) even though her terminology is somewhat unorthodox.

2. Supererogation and Dual-Role Views of Reasons

(that is, neither save Adam nor go to the restaurant), we can directly state a formal paradox based on *Life or Promise*.

1. $Can(W|W \vee A)$ (intuition)
2. $Can(A|A \vee P)$ (intuition)
3. For all X, Y, Z , if $Can(X|X \vee Y)$ and $Can(Y|Y \vee Z)$, then $Can(X|X \vee Z)$
(intuitive principle)
4. Not $Can(W|W \vee P)$ (intuition)

The intuitions about the case seem quite plausible if one accepts supererogation as a phenomenon at all. The principle used here, that permission is transitive in pairwise choices, seems very plausible.³¹

We can also easily formalize the case. We abbreviate W – walking away – as $\neg P \wedge \neg A$, and let L , like before, stand for the fact that you lose your legs. Then the context $\Delta_4 = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$ can be expressed as follows, $\mathcal{W} = \{A \rightarrow L, P \rightarrow \neg A, A \rightarrow \neg P\}$, $\mathcal{R} = \{!P, !A\}$, $\mathcal{J} = \mathcal{R} \cup \{!\neg L\}$ with $!P < !A$. We treat the reason grounded in your promise as a requiring reason as it clearly seems to have requiring force upon you, and whether you follow it does not seem up to you in the same way as keeping your legs is. Nonetheless, it is clearly a weaker reason than the reason to save Adam.

We can derive both $Can(W|W \vee A)$ and $Can(A|A \vee P)$ straight away. But, as intended, $Can(W|W \vee P)$ cannot be derived. That means that here, too, we must deny the principle. Permission cannot be transitive. According to our model, it is not. In fact, Δ_4 itself serves as a counterexample.

Does the model also help us to explain the intuitive pull of the transitivity principle? We might be tempted to diagnose that here, too, the pull stems from the fact that two related principles hold. In our model, for the other two modalities, we do get transitivity under pairwise choices.

Proposition 2.4.2. *For any context Δ ,*

- *if $\Delta \vdash Ought(X|X \vee Y)$ and $\Delta \vdash Ought(Y|Y \vee Z)$ then $\Delta \vdash Ought(X|X \vee Z)$.*

³¹There are other supposed cases of transitivity violation, for example the *Car Case* from (Dorsey, 2013). There might be significant differences (as implicitly claimed by (Tucker, forthcominga)) or they might be more or less the same (as claimed by (Muñoz, 2020)). We focus on Kamm here as it is probably the most discussed transitivity violation and will be most interesting to discuss in the context of our formal model.

- if $\Delta \vdash \text{Must}(X|X \vee Y)$ and $\Delta \vdash \text{Must}(Y|Y \vee Z)$ then $\Delta \vdash \text{Must}(X|X \vee Z)$.

The proofs for these two statements are not straightforward and can be found in Appendix A.2.

We could conclude here and claim requiring and justifying reasons can account for all three paradoxes of supererogation, and see this as a point in favour of our model. But this is too quick. In Section 2.5, we will show that the case motivating the Classic Paradox of Supererogation, as presented here, indeed cannot be adequately modeled without justifying reasons. Kamm's *Life or Promise* case on the other hand, *can* be modeled without resorting to justifying reasons. We will spend the remainder of this section showing and discussing this fact.

Let us consider Kamm's case again with just a single type of reason, that is $\Delta_3^* = (\mathcal{W}, \mathcal{R}, <)$ with $\mathcal{W} = \{A \rightarrow L, P \rightarrow \neg A, A \rightarrow \neg P\}$, $\mathcal{R} = \{!P, !A, !\neg L\}$ and $!P < !A$. That is, this (single-type) context is identical to Δ_4 , except that instead of the set $\mathcal{J}$ of justifying (and requiring) reasons, we now only have the single set $\mathcal{R}$.

Observation 2.4.3. *Letting W abbreviate $\neg A \wedge \neg P$, We can see that*

- $\Delta_3^* \vdash \text{Can}^*(W|W \vee A)$
- $\Delta_3^* \vdash \text{Can}^*(A|A \vee P)$
- *but not* $\Delta_3^* \vdash \text{Can}^*(W|W \vee P)$
- *and thus not:* If $\text{Can}^*(X|X \vee Y)$ and $\text{Can}^*(Y|Y \vee Z)$, then $\text{Can}^*(X|X \vee Z)$ for any X, Y, Z

The proofs of these statements are quite straightforward and can be found in Appendix A.2. We can thus see that even in a context with just one type of reason, all statements save for the transitivity principle hold, but the principle itself does not. What's more, just like before, we can explain (in this single-type context) what makes the principle so intuitive.³²

Namely, that

Proposition 2.4.4. *For any single-type context Δ^* ,*

- *if* $\Delta^* \vdash \text{Ought}^*(X|X \vee Y)$ *and* $\Delta \vdash \text{Ought}^*(Y|Y \vee Z)$ *then* $\Delta \vdash \text{Ought}^*(X|X \vee Z)$.

³²Note though, that in the simple way to get oughts, as introduced in footnote 6, permissions are indeed transitive under pairwise choices.

2. Supererogation and Dual-Role Views of Reasons

The proof for this proposition can be found in Appendix A.2.

But this means, something else is going on here. What's doing the heavy lifting here is not the justifying/requiring distinction, but the construction of the Brewka scenarios. (Muñoz, 2021) proposes that in order to make sense of this purported transitivity violation, we should accept that normativity is essentially two-dimensional – that is, that we need to distinguish between justifying and requiring reasons. The alternative, he claims, would be to accept a strange kind of menu sensitivity – the thesis that reasons can change their strength depending on what actions one can perform in some situation. We can see that this is not the case here. The single-type solution to the paradox is neither menu sensitive, nor does it make use of justifying and requiring reasons. We think that Muñoz' mistake is assuming what we have called the *kitchen scale model* in Section 1.4.2. Every reason adds some weight to a single scale that determines choice-worthiness (which indeed forces transitivity). But this is not how reasons behave in our model. In the pairwise choice, the absence or presence of some reasons can influence the relative ranking of actions.

Is our model then indeed one where we have to accept some strange kind of menu sensitivity? We know that some reasons can become *inactive* – they are not included in any Brewka scenario any more – in a conditional context (here, that is a context where we can only choose between two actions). Does this rise to the kind of menu sensitivity that Muñoz worries about? We don't think so. For one, of course, the partial order $<$ on the reasons remains unchanged. Additionally, the fact that some reasons get excluded is simply how any obligation (conditional or not) is determined by the Brewka construction we base our model on. So either every such model applied to the practical domain exhibits this strange menu sensitivity, or there is a third alternative to *two-dimensional* or menu sensitive approaches – those based on Brewka's construction.

The thesis that we need an additional dimension of normativity (or menu relativity) is thus tied to a specific way of deriving permissions from reasons: one that treats reasons as essentially contributing a certain weight to an option. We will discuss the central concept behind this view in Chapter 3. But for now, we can confidently say that it is incorrect that Kamm's Intransitivity Paradox can only be accounted for by using both justifying and requiring reasons. The way of deriving requirements (or oughts) from a single type of reason inspired by Brewka's work is enough to fully account for the puzzle surrounding Kamm's case. Only when combined with additional assumptions does the need for a second dimension of normativity arrive in Kamm's case.

We have seen that our model can give an answer to the *Classic Paradox of Supererogation*: why should we not always simply do what's best. We have also seen that two other paradoxes – The All or Nothing Paradox and the Intransitivity Paradox – do not lead to any problems for the model. We have furthermore suggested, why the cases seem to engender paradoxes. Lastly, contra Muñoz, we have shown that Kamm's paradox does not necessarily arise out of a disregard for the distinction between requiring and

justifying reasons and only additional assumptions make them strictly necessary to solve the paradox. Proponents of the existence of supererogation thus have further options in how to respond to the challenges raised by the puzzles.

2.5. Modeling the Cases Without Justifying Reasons

We have claimed that introducing justifying reasons is not necessary to account for Kamm's case. The same cannot be said about Horton's All or Nothing case, as the principle makes use of the stronger modality *Must*(·). But what about the first case that led to a paradox, *Save One or Two*, the case where you could save one person at no costs, but saving two would cost you your legs?

In this section, we will discuss if and how we could model the case to capture the intuitions while using only a single type of reason and argue that, ultimately, all such attempts fail.

Let's recall how we've modeled the case formally. We introduced the dual-role context $\Delta_1 = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$, where $\mathcal{W} = \{A \wedge B \rightarrow L\}$, $\mathcal{R} = \{!A, !B\}$, $\mathcal{J} = \mathcal{R} \cup \{!\neg L\}$. For the order we included $!A < !\neg L$ and $!B < !\neg L$.

We have already seen that if we transpose this context to the single-type context Δ_2^* , where we do not distinguish between requiring and justifying reasons, $Can^*(A \wedge B)$ cannot be derived.

A possible response is that our choice of $<$ was illegitimate. So let's explore what other possibilities we have for the choice of $<$. If, for the context Δ_4^* , we were to reverse the order, setting $!\neg L < !A$ and $!\neg L < !B$, we would get $\Delta_4^* \vdash Can^*(A \wedge B)$ (in fact, we would get the stronger $\Delta_4^* \vdash Ought^*(A \wedge B)$, but no longer get either of the intuitive results $\Delta_4^* \vdash Can^*(A \wedge \neg L)$ or $\Delta_4^* \vdash (B \wedge \neg L)$. So this attempt also fails. Perhaps, we should not postulate any order holding between the reasons.

This leads us to $\Delta_5^* = (\mathcal{W}, \{!A, !B, \neg L\}, <)$ where $\mathcal{W}$ is as before, but $<$ is empty. Here we get the intuitive results $\Delta_5^* \vdash Can^*(A \wedge \neg L)$, $\Delta_5^* \vdash Can^*(B \wedge \neg L)$ and even $\Delta_5^* \vdash Can^*(A \wedge B)$. But we no longer can say anything about the fact that there is something *really good* about saving both Adam and Benoît, as $Ought^*(A \wedge B)$ no longer follows from Δ_5^* .

So as long as we have modeled the cases correctly, that is, we have represented the reasons and $\mathcal{W}$ adequately, then the single-type-model simply cannot account for all supererogation phenomena correctly. Are there other ways to represent the reasons?

2. Supererogation and Dual-Role Views of Reasons

Maybe, what we have done is an inadequate formalization of the situation where we only look at individual reasons, but we know that the model cannot accommodate how the strength of multiple reasons *adds up* (we will discuss this phenomenon of the aggregation or accrual of reasons in more detail in Section 2.7). Let us thus set up a different formalization, where instead of having individual reasons to save Adam and Benoît, we have two reasons, $!S1$ and $!S2$ standing, respectively, for the fact that you have a reason to save one person and a reason to save two people. We then set $\mathcal{W} = \{S2 \rightarrow S1, S2 \rightarrow L\}$ and $\mathcal{R} = \{!S1, !S2, !\neg L\}$. For the scenario to be plausible, we also want $!S1 < !S2$, as clearly, whatever else may be going on, saving two people is more important than saving one person.³³

And indeed, there is some partial order that respects this and gives us the results we want, namely the smallest partial order such that $S2 > S1$.³⁴ In this context, Δ_6^* , we get the intuitive result that $\Delta_6^* \vdash Can^*(S2)$, $\Delta_6^* \vdash Ought^*(S1)$ and $\Delta_6^* \vdash Can^*(\neg L)$.

So, someone defending the view that we don't need a distinction between justifying and requiring reasons to accurately model at least some paradoxes surrounding supererogation has options here. However, we don't believe they are good ones.

For one, unsurprisingly, we do not get a distinction between the two modalities that give cases of supererogation their bite. We *must* save one person, but we *ought to* save both.³⁵

The bigger problem, we think, however, is that formalizing the case using $!S1$ and $!S2$ as the reasons, is a somewhat ad-hoc choice here. For one, clearly, the normative force of the reasons comes from Adam's and Benoît's respective well-being, and so we would expect a faithful formalization of the case to include that. Second, having a single reason in favour of saving both their lives is not as natural as it first appears. We do not make use of the accrual of reasons anywhere in our cases, and here, too, we do not have a situation where two weaker reasons together (that is $!S2$) defeat the reason given by $!\neg L$, while each reason individually (that is $!S1$) does not. This leads us to conclude that formalizing the case in this way does not have any independent motivation apart from that it solves the puzzle.

So to take stock: Either we have to give up on modeling the *Save One or Two* case

³³This is regardless of whether one believes “the numbers should count” (see (Taurek, 1977)) or not, as we are here dealing with the choice between saving one person or saving the same person in addition to one other person. So this order is the only plausible one we can give.

³⁴Additionally, the empty partial order gives us the same results, but is less intuitive.

³⁵An interesting, if ultimately unsuccessful, idea can be found in (Mullins, 2021). Mullins takes up an idea by (Scanlon, 2014), that the difference between requiring and justifying reasons depends on their strength, where requiring reasons are stronger and justifying reasons are weaker. This approach allows Mullins to define two deontic modalities with a single type of reason. However, as can be seen in many of our examples, the idea that justifying reasons are weaker than requiring reasons fails to provide the desired results in those cases.

naturally, or we need to distinguish between justifying and requiring reasons. We think that in such a situation the choice is easy, and we should go with the one that is independently motivated and intuitive: distinguishing between justifying and requiring reasons.

2.6. Related Work

In this section, we will compare our model to existing approaches in the literature. In Section 2.6.1, we will look at non-formal accounts of justifying and requiring reasons. The first account we look at is, we think, the account that most closely resembles ours – (Muñoz and Pummer, 2022) and (Pummer, 2023). Next we will look at a claim our model makes: all requiring reasons are also justifying reasons, and compare it with differing approaches. Ultimately, we will argue, at least for a model like ours, this is indeed the only sensible way: requiring reasons are justifying reasons. In Section 2.6.2, we will look at formal work. First, we will look at the models of supererogation by Mullins (2021) and the *Doing Well Enough Framework* by McNamara (2022b, 2025). Next, we look at (Hansen, 2014) who uses a similar approach to ours to model explicit permissions.

2.6.1. Non-Formal Approaches

The Account of Pummer & Muñoz

The closest overlap with the way our model derives obligations is found in the non-formal account of (Muñoz and Pummer, 2022) and (Pummer, 2023). Here Muñoz and Pummer offer both an account of the two roles reasons play and a way to generate conditional obligations. First, let’s see how (Pummer, 2023) characterizes what justifying and requiring reasons do.³⁶

An act is required – wrong not to do – when there is most requiring reason overall to do it and no sufficiently strong [justifying] reason not to do it. A [justifying] reason is a kind of “defeater”: when it’s sufficiently strong, it prevents the balance of requiring reasons from making an act wrong. (Pummer, 2023, p. 10-11)

Even though this is couched in terms of the balance of reasons, the overall idea is similar to the one we present here. Justifying reasons can *defeat* requiring reasons, by

³⁶Pummer uses the term “permitting reasons”, in the following quote, we replaced it by the term “justifying” reasons, like used in (Muñoz and Pummer, 2022).

2. Supererogation and Dual-Role Views of Reasons

taking away their power to make an act required.³⁷ In our model, this works through the selection of Brewka Scenarios. An act is only required if it's a consequence of (the conclusions of) all Brewka Scenarios. A justifying reason's contribution is that it can replace some requiring reasons in some Brewka Scenarios, and so prevent some consequence of (the conclusions of) those requiring reasons to become required. This is how justifying reasons can defeat requiring reasons in our model. Muñoz and Pummer do not offer such a detailed account of how the defeat happens. Instead, they rely on the intuition of a model of weighing reasons – the topic of Chapter 3. There is another important distinction to keep in mind. Like (Tucker, forthcominga), (Pummer, 2023) treats reasons as contrastive.³⁸ Our model, on the other hand, does not need to make such a concession. The most important similarity, however, is probably in how conditional modalities are treated. To our knowledge, Pummer and Muñoz are the only ones addressing conditional requirements, obligations, and permissions explicitly in the context of dual-role views on reasons.

To illustrate this, we look at the case of conditional permission $PE(\cdot|\cdot)$ as defined by Muñoz and Pummer:

$PE(B|\neg A)$ iff for any alternative C that is compatible with $\neg A$, the requiring reason to do C (rather than B) does not outweigh the justifying reason to do B (rather than C). (Muñoz and Pummer, 2022, p. 1438)

Again, this is couched in both contrastive and weighing reasons terminology, but the overall strategy is quite similar to ours. Both our model and Muñoz and Pummer throw out reasons incompatible with the condition ($\neg A$ in their definition). They do it by excluding alternative actions that are incompatible with the condition (and thus disregard reasons that speak in favour of that action), we do it by not counting reasons conflicting with the condition as *active* and thus prevent their inclusion in any Brewka scenario. What is also similar is that justifying reasons can prevent an obligation to come into effect. In Muñoz and Pummer's case in terms of *outweighing*, in our case, through preventing some conclusion from being derivable from every Brewka scenario. So, while the way deontic modalities are derived differs – models in the style we presented, including that of (Horty, 2012), do not allow for any kind of (out)weighing of reasons – the overall approach is quite similar.

³⁷We will see a formal model that is closer to this balance model in Chapter 3.

³⁸A reason is contrastive if it is only a reason (for/against some action) relative to some alternative action. The term comes from (Snedegar, 2017) who is also the most prominent defender of such a view.

Do Requiring Reasons Justify?

The two most prominent authors advocating for a distinction between justifying and requiring reasons are (Gert, 2007, 2004, 2011, 2014) and (Greenspan, 2005, 2007). Recently, this distinction has been defended by (Tucker, forthcominga, 2021).³⁹ In this section, we will briefly highlight the relevant differences these proposals have with ours, while focusing particularly on the relation between justifying and requiring reasons.

Our model presumes that every requiring reason is also a justifying reason. In fact, this idea is hard-coded in our notion of dual-role contexts. (Muñoz and Pummer, 2022) present this as the default idea, but there are some detractors.

(Greenspan, 2005, 2007) seems, like us, to treat requiring reasons – “negative” reasons in her terminology – as essentially encompassing a justifying – “positive” – element as well. In Greenspan’s view, roughly, what rationality requires of one is having an answer to criticism. Requiring reasons give rise to criticism of alternative actions, while justifying reasons can give an answer. That means, that, like our model does, Greenspan treats purely justifying reasons as essentially optional elements of one’s decision-making.

On the other hand, (Gert, 2007) appeals to two entirely distinct kinds of “normative strengths”. While he does seem to accept that in a wide variety of cases, requiring reasons also justify, arguably, he would see the fact that requiring reasons *always* justify as the same kind of “cosmic coincidence” (see (Gert, 2007, p. 543)) as if justifying and requiring strengths of reasons always exactly correspond to one another, thus making the distinction between the two kinds of roles reasons can play disappear.

In a similar vein, (Tucker, forthcominga), claims that reasons come with two kinds of strengths, corresponding to a justifying and requiring dimension. Tucker explicitly endorses the view that in some cases the requiring strength can be larger than the justifying strength.⁴⁰ In extreme cases, this means that some reasons could even be purely requiring, without exerting any justifying force. If we adopt Tucker’s view and translate it to our model this means that we would need to allow for purely requiring reasons.

Let us briefly consider what it would mean for our model to drop the requirement that requiring reasons are also justifying reasons. In effect, it would lead us to posit three disjoint sets of reasons, $\mathcal{J}$, the set of purely justifying reasons, $\mathcal{R}$, the set of purely requiring reasons and $\mathcal{M}$, the set of and reasons that are both requiring and justifying.

³⁹This distinction has of course also been defended by (Muñoz, 2021) and (Muñoz and Pummer, 2022; Pummer, 2023) who we discussed above.

⁴⁰Tucker does this to accommodate yet another puzzle about a supposed intransitivity violation; Dorsey’s Car Case (see (Dorsey, 2016)). We don’t believe Dorsey’s car case leads to a genuine paradox, as the claims one needs to accept about the case strike us as neither particularly intuitive nor independently plausible, but we will not defend that thesis here.

2. Supererogation and Dual-Role Views of Reasons

This means that our model would now be $(\mathcal{W}, \mathcal{J}, \mathcal{R}, \mathcal{M}, <)$ with the relation $<$ being defined on the union of the three sets of reasons.

We conjecture that the only way to derive oughts and requirements from this model that still satisfies principles like the ones introduced in Section 2.3 while making sense of the cases we presented would work as follows:

We again build stable scenarios, which are sets $\mathcal{D}$ such that $\mathcal{R} \cup \mathcal{M} \subseteq \mathcal{D} \subseteq \mathcal{R} \cup \mathcal{M} \cup \mathcal{J}$ and then, once more, use the Brewka scenarios based on $\Delta^{\mathcal{D}}$ to determine requirements and permissions. The only sensible way to derive oughts, it strikes us, is to use the stable scenario $\mathcal{R} \cup \mathcal{M}$ as the basis, just like we used $\mathcal{R}$ as the basis in our original model.

But what we have, in effect, is just a new description of our old model, where $\mathcal{R} \cup \mathcal{M}$ takes the place of what was originally just called $\mathcal{R}$ and the set of all reasons takes the place of $\mathcal{J}$. If this line of reasoning is correct, what we are left with is just a terminological difference and, at least when viewed through the lens of our models, requiring reasons always also justify.

Lastly, we will briefly mention a thesis by (Scanlon, 2014). Scanlon claims that we can distinguish justifying and requiring reasons by their relative strength. The requiring force of requiring reasons, on this account, roughly, comes from the fact that they are decidedly stronger than justifying reasons. We do not think this is the right way to characterize them. In fact, we have seen in Section 2.5 that in order to model the cases, particularly the *Save One or Two* scenario, in our model, we cannot allow for the justifying reasons to be always weaker than the requiring reasons. So, if the model of reasons we presented is an appropriate tool for modeling how reasons generate oughts and requirements at all, we must conclude that Scanlon's suggestion fails.⁴¹

In this section, we have seen that our approach is in some respects quite close to some non-formal approaches found in the literature. The biggest difference, we have claimed, is that our model does not *weigh* the different types of reasons. We have argued that our model provides support for the thesis that every requiring reason always justifies – contra Gert and Tucker. After having compared our model to the non-formal approaches in the literature, we will compare it to the formal side next.

⁴¹(Mullins, 2021) discusses and ultimately rejects a formal version of this Scanlonian approach to justifying and requiring reasons. We will not discuss his proposal in detail here. We do not mean to claim that it is inherently impossible for such an approach to work, but it cannot make use of the tools we have presented here.

2.6.2. Formal Approaches

On the formal side, there are two other proposals that also model oughts and requirements: Mullins' (2021) model and McNamara's Doing Well Enough Framework (see (McNamara, 1996a,b, 2011, 2022a,b)) Additionally, we note similarities with work on modeling explicit permissions in deontic logic with our model. Arguably, the most closely related formal structures can be found in (Hansen, 2014), which we discuss at the end of this section.

Mullins

Mullins' (2021) model is based on Horty's (2012) take on default logic and, like us, introduces a distinction between requirements and oughts. In the paper, Mullins offers both a model based on the strength of reasons and one based on dual-role views of reasons. We focus our discussion on the latter model.

Unlike us, Mullins relies on the simple account to defeat (see footnote 6). Additionally, whereas we determine oughts (i. e. *Ought*($\cdot$) formulas) on the basis of requiring reasons and musts (i. e. *Must*($\cdot$) formulas) through the interplay between justifying and requiring reasons, Mullins takes the exact opposite approach. He obtains requirements on the basis of requiring reasons alone whereas oughts are then determined, roughly, based on the reasons that are compatible with the derivation of *Must*($\cdot$) formulas.

Explaining his strategy, Mullins writes:

We first identify our undefeated requiring reasons, in order to determine what is required or impermissible. Oughts are then supported by our best justifying reasons, provided the consequences of their conclusions are consistent with some maximal subset of requiring reasons (Mullins, 2021, p. 586).

We will explain his approach in detail. His form of (dual-role) contexts look basically identical to ours, except that in Mullins' case, we do not have the additional assumption that $\mathcal{R} \subseteq \mathcal{J}$, that is, every requiring reason is not necessarily a justifying reason. The partial order $<$ in Mullins' model ranges over all reasons in the context (we also include the set of background factors $\mathcal{W}$ explicitly in a context, Mullins does not).

We will use the same notation for Mullins' requirements, oughts and permissions, but use $\vdash^M$ to indicate that a formula follows from a context in Mullins' system. We will also adapt the presentation of Mullins' definition slightly to make them correspond closer to our definition of dual-role contexts above.

2. Supererogation and Dual-Role Views of Reasons

Mullins then defines (Mullins, 2021, p. 585):

Definition 2.6.1. *For some context $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$*

- $\Delta \vdash^M \text{Must}(X)$ *iff $\mathcal{W} \cup M \vdash X$ for every maximal consistent (with $\mathcal{W}$) subset M of $\text{Conclusions}(\text{Undeclared}(\mathcal{R}))$.*
- $\Delta \vdash^M \text{Can}^M(X)$ *iff it is not the case that $\Delta^M \vdash \text{Must}(\neg X)$.*

In order to define Oughts, Mullins defines:

Definition 2.6.2.

Given a context $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$, M is requirement-consistent iff $M \cup M'$ is consistent (with $\mathcal{W}$) for some consistent subset M' of $\text{Conclusions}(\text{Undeclared}(\mathcal{R}))$.

with this, Mullins defines Oughts as:

Definition 2.6.3. $\Delta \vdash^M \text{Ought}(X)$ *iff $M \cup \mathcal{W} \vdash X$ for some maximal requirement consistent subset M of $\text{Conclusion}(\text{Undeclared}(\mathcal{J}))$.*

To see Mullins' model at work, we revisit the familiar Save One or Two scenario. We captured it in the context $\Delta_1 = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$ where $\mathcal{W} = \{(A \wedge B) \rightarrow L\}$, $\mathcal{R} = \{!A, !B\}$, $\mathcal{J} = \{!A, !B, !\neg L\}$, and $<$ empty. To determine which *Must*-formulas follow from it, we are to look at what follows from the subsets of undefeated requiring reasons – that is, the subsets of $\text{Conclusion}(\mathcal{R}) = \{A, B\}$ – that are maximal and consistent with $\mathcal{W}$. This, however, gives us the counterintuitive result that both $\text{Must}(A \wedge B)$ and $\text{Must}(L)$ follow from Δ_1 : you have to save Adam and Benoît, and lose your legs. Perhaps the fault lies in the modeling of the case, and it has to be described by a different context. The most natural alternative that suggests itself is $\Delta_5 = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$ where $\mathcal{W} = \{(A \wedge B) \rightarrow L\}$, $\mathcal{R} = \emptyset$, $\mathcal{J} = \{!A, !B, !\neg L\}$, and $<$ is empty.⁴² Even barring the counterintuitive implication that the fact that Adam and Benoît are in danger doesn't

⁴²Another possible candidate is the context $\Delta_6 = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$ where $\mathcal{R} = \{!(A \vee B)\}$ and the rest is like in Δ_5 . But while this secures the intuitive results that both $\text{Must}(A \vee B)$ and $\text{Ought}(A \wedge B)$ are derivable, there's good reason to be dissatisfied with this context. Most importantly, the inclusion of a *disjunctive* requiring reason looks terribly ad hoc, since it amounts to hard-coding the desired intuition. Also, $\text{Ought}(\neg L)$ follows from Δ_6 , just like it does in the case of Δ_5 . Perhaps the problem is that we leave the ordering relation $<$ empty. Since justifying reasons can defeat requiring reasons in Mullins' model (but not ours), does this stack the deck against him? This doesn't seem to be the case. There appears to be no ordering that delivers the desired result. Setting $!A, !B < !\neg L$ fails to deliver the intuitive $\text{Must}(A \vee B)$, while adding $!(A \wedge B)$ to $\mathcal{R}$ and setting $!(A \wedge B) < !\neg L$ does not help either.

exert requiring normative force, we don't get a good match with our intuitions. First, $Must(A \vee B)$ doesn't follow from Δ_5 . Second, while $Ought(A \wedge B)$ follows from Δ_5 , so does $Ought(\neg L)$, suggesting that not losing your legs is optimal.⁴³

We take all of this as evidence that Mullins' model has serious trouble accommodating cases of supererogation, specifically in modeling the Save One or Two scenario. We think that the problem is that Mullins' model has the order of explanation wrong. Justifying reasons have too weak an impact on requirements: Only by outright defeating requiring reasons do they play any role at all. This position is in line with a non-formal proposal by (Snedegar, 2016). But we think that, at least for models like Mullins' (or ours) such a position is not tenable.

McNamara

The oldest well-developed formal account of supererogation is McNamara's Doing Well Enough Framework, published in a series of papers (see (McNamara, 1996a,b, 2011)). Here, we will focus primarily on two recent additions that allow us to compare it with our approach directly. First (McNamara, 2022a), allowing the framework to express conditional obligations and second, (McNamara, 2025), McNamara's recent attempts to account for puzzles like the All or Nothing Puzzle and supererogatory holes in general. We will discuss these two additions as a whole, even though they do not appear in the same paper together.

McNamara's Doing-Well-Enough Framework is a modal logic where a frame $(W, A, \leq)$ consists of a set of possible worlds – which we will here throughout assume to be finite – where the worlds are ranked by their goodness via $\leq$. The set A partitions the set of worlds into “acceptable” and “unacceptable worlds”.⁴⁴ Acceptable worlds, roughly, are those worlds, where the agent performs a permissible action; unacceptable worlds are, roughly, those where she does not. In order to make sense of the “goodness” of worlds, $\leq$ is assumed to be transitive, reflexive, and connected.⁴⁵ McNamara uses three operators that we think correspond to our $Must(\cdot)$, $Ought(\cdot)$ and $Can(\cdot)$ formulas. These are $OB(\cdot)$, standing for “it is overridingly obligatory”, $PE(\cdot)$, standing for “it is permissible”, and $MA^P(\cdot)$ standing for “the best one can permissibly do”. $OB(\phi)$ follows just in case ϕ is true in every acceptable world and $PE(\cdot)$ is its dual operator. $MA^P(\phi)$, on the other hand, only follows if ϕ is true in every *best* acceptable world. McNamara goes on to define various other operators that have no correspondence in our model,

⁴³Mullins' strategy uses the *conflict account* (picking a maximal consistent set) in determining which oughts follow from a context. A natural idea is to substitute it with the *disjunctive account* (deriving an ought only if it follows from *all* maximal consistent sets). Unfortunately, this move doesn't solve the problem: $Ought(\neg L)$ no longer follows from the context, but neither does $Ought(A \wedge B)$.

⁴⁴Strictly speaking, both $\leq$ and A are defined relative to each world in W . While this is important for the semantics proper, we leave this complication aside for the sake of a simpler exposition.

⁴⁵McNamara also discusses weaker versions, but we will ignore these here.

2. Supererogation and Dual-Role Views of Reasons

like, for example, the idea of the least one can permissibly do without doing something one must not do. If we add another constraint we are back to the standard account of supererogation mentioned in section 2.4: Every world in A , the set of acceptable worlds, is better than every other world; that is, there is a threshold of goodness that separates acceptable and unacceptable worlds. If we hold on to this constraint, we can derive $PE(X) \wedge OB(\neg Y) \rightarrow OB(X|X \vee Y)$, which corresponds to the principle we rejected in Section 2.4 when discussing the All or Nothing Paradox.⁴⁶ This means, that given this constraint, the Doing Well Enough Framework cannot resolve the paradox. Similarly, we can derive a version of the transitivity of permission and thus cannot easily accommodate Kamm’s Intransitivity Paradox.

The first part, the All or Nothing Problem, leads McNamara to abandon this constraint. And indeed, once we give up on the notion of a threshold, the principle can no longer be derived (see (McNamara, 2025)). But this leaves an explanatory gap. Where does the partition into (un)acceptable worlds come from? The threshold presents a nice story here: at some point, you’re doing *well enough*. What can we say about the other option? One answer is, of course, that it is not the job of the logic to say anything about how models are constructed, but this is too hasty a response. In our model, we rely (besides the semantics) on a single input: Justifying and requiring reasons and their relative strength. Adding the constraint makes the Doing Well Enough model equally intuitive: We have an order on worlds and a threshold, both to be supplied from first-order ethics. With this, we can define when an action is permitted, required or ought to be done. Without the threshold constraint, however, we need another thing: for each world, we must at the outset tell the logic whether it is acceptable or not. The logic itself then loses explanatory power. It can no longer explain what makes an act acceptable by simply deferring to the idea that first-order ethics provides a threshold of goodness. In fact, it no longer has any answer to give; What makes a world acceptable now falls outside the purview of what the logic can answer. Our model, on the other hand, can give a clear answer: An act is permissible when all requiring reasons not to do it are defeated by a justifying reason.

Hansen

So far, we have discussed related work that, like us, intends to model the distinction between *requirements*, *oughts*, and *permissions*. That is, we have focused on work whose content is similar to ours. In addition to this, there is related work where the content is different, but the formal tools used are quite similar. This is work on explicit permission in deontic logic. Explicit permission is best understood in legal or quasi-legal contexts. Some authority has given a set of orders that should be obeyed, as well as a set of explicit permissions to do certain actions. If there is no conflict between the orders and

⁴⁶Here we make use of (McNamara, 2022a), who introduced conditional operators to the Doing Well Enough Framework.

the permissions, we have to obey the orders, but can otherwise do as we please, but what if some permission explicitly allows something that an order prohibits? How are we to decide what we are allowed to do then? These questions have been explored in different frameworks by different authors (see for example (Makinson and van der Torre, 2003; Boella and van der Torre, 2008; Stolpe, 2010; Hansen, 2014; Olszewski et al., 2023)). Like in our model, we are dealing with two types of normative features, orders, and permissions in one case and requiring and justifying reasons in the other. In fact, on the formal level, we can find some overlaps between these two approaches, the biggest of which is arguably (Hansen, 2014), which we discuss next.

(Hansen, 2014) introduces three different semantics. We focus on the third he introduces, as it is both the closest to our model and the most general. We will slightly adapt the notation to bring the similarities into sharp relief.

With propositional logic as a backdrop, Hansen starts with two (disjoint) sets of “imperatives” I and “licences” L , collected in a tuple (I, L) . Imperatives stand for orders one is given, and licences are explicit permissions. These two sets, like in our model, are made up of $!$ -formulas, where, again, the conclusion of a formula $!A$ is simply A .⁴⁷

Hansen then introduces a strict partial order $<$ on $I \cup L$, much like we do on the set of all reasons $\mathcal{J}$. Hansen also includes a set of propositional formulas that aim to model the facts that hold true in some situation. We model the same thing, and so we are going to use the same notation here, namely $\mathcal{W}$.⁴⁸ While we include $\mathcal{W}$ in our model, Hansen instead uses a modality that conditionalizes on $\mathcal{W}$. This is a mere notational variant and we will ignore this complication throughout, treating $\mathcal{W}$ as a part of the model in Hansen’s case. Hence, we define a Hansen model as a tuple $\Delta^H = (\mathcal{W}, I, L, <)$ analogous to our (dual-role) contexts.

Next, Hansen defines a deontic modality $O(\cdot)$ that represents what is (still) obligatory in a situation $\mathcal{W}$, where both imperatives I and licences L are in effect. While some of the semantics Hansen uses will still be quite similar to ours – in fact we aim to present his semantics in as similar a manner to ours as possible – this is where we will start to see the first differences. The first of which is that Hansen does not use the ordinary notion of consistency on the set $\text{Conclusion}(I \cup L)$, but instead uses what he calls “Von Wright Consistency” (after (von Wright, 1963)).

Definition 2.6.4 (Von Wright Consistency). *Given a tuple of imperatives (I, L) and a context $\mathcal{W}$, we say that (I, L) is von Wright consistent with $\mathcal{W}$ iff $\text{Conclusion}(I) \not\models \neg \mathcal{W}$*

⁴⁷There is some notational difference here. First, Hansen uses $!$ formulas only for imperatives and uses formulas prefixed by l for licences. So a licence could have the form $l(A)$. Furthermore, to be more general, Hansen allows for imperative and licences to have an arbitrary Conclusion, that is given by a specific function f . However, we will skip over this complication here, as it is immaterial to the comparison we pursue here.

⁴⁸Hansen uses the letter C .

2. Supererogation and Dual-Role Views of Reasons

and for every $l \in \{k \in L : \text{Conclusion}(k) \not\models \neg \mathcal{W}\}$, $\text{Conclusion}(I \cup \{l\}) \not\models \neg \mathcal{W}$.

That means, that in order for (the conclusions) of a tuple (I, L) to count as consistent, it is enough if each licence is individually consistent with the imperatives (and the facts of the situation).⁴⁹ The licences themselves need not be consistent with each other. The intuition behind this definition is that if one has the permission to do inconsistent things, that is okay, both will restrict, as we will see, one's obligations, but they do not interfere with each other. We use a similar idea, but in our model, what does the work is the notion of a stable scenario, where one is free to include whatever justifying reasons one sees fit. Hansen does not have a similar notion, and this is the second major difference. To see Hansen's approach in action, we can again model the notation very closely to ours. This gives us, first, an analogue of our active reasons.

Definition 2.6.5 (Adding a Licence or Imperative). *Given a tuple of imperatives and licences (I, L) a set of facts $\mathcal{W}$ and a tuple (I', L') where both $I' \subseteq I$ and $L' \subseteq L$, where $r \in I \cup L$, we define $(I', L') + r$ as $(I' \cup \{r\}, L')$ if $r \in L$ and as $(I', L' \cup \{r\})$ otherwise.*

Definition 2.6.6 (Active Imperatives and Licences). *Given a Hansen Model $\Delta^H = (\mathcal{W}, I, L, <)$ and a tuple (I', L') where both $I' \subseteq I$ and $L' \subseteq L$, let*

$$\text{Active}_{\Delta^H}(I', L') = \{r \in I \cup L : (I', L') + r \text{ is von Wright consistent with } \mathcal{W} \text{ and } r \notin I' \cup L'\}.$$

We analogously define this operation for sets of imperatives or licenses.

Hansen's next step is very similar to ours. He uses the construction of (Brewka, 1994) to use the partial order on imperatives and licences to construct an analogue of our Brewka Scenarios. Again, we make the notation closely match ours.

That is, given some Hansen model $\Delta^H = (\mathcal{W}, I, L, <)$ and a total extension $\prec$ of $<$, he defines, much like we did:

Definition 2.6.7 (Hansen-Brewka-Tuples). *We define the Hansen-Brewka tuple (I', L') of a Hansen model $\Delta^H = (\mathcal{W}, I, L, \prec)$ as $(I', L') = (\bigcup_{i \geq 0} I'_i, \bigcup_{i \geq 0} L'_i)$, where the sequence $(I'_0, L'_0), (I'_1, L'_1), (I'_2, L'_2), \dots$ is defined as follows:*

$$\begin{aligned} (I'_0, L'_0) &= (\emptyset, \emptyset), \\ (I'_{i+1}, L'_{i+1}) &= \begin{cases} (I'_i, L'_i) & \text{if } \text{Active}_{\Delta^H}(I'_i, L'_i) = \emptyset \\ (I'_i, L'_i) + \text{Max}_{\prec}(\text{Active}_{\Delta^H}((I'_i, L'_i))) & \text{otherwise}^{50} \end{cases} \end{aligned}$$

⁴⁹At the end of the paper, Hansen briefly mentions a stronger version of consistency, where not each individual licence has to be consistent with the set of imperatives, but maximal subsets of licences need to be. This is closer to what we use in our model, but we will not discuss it in detail here.

With this definition in place, we can define Hansen’s main operator $O(\cdot)$, which we will read here (to keep it distinct from our ought and requirement operator) as “There is an obligation to $\cdot$ ”.

Definition 2.6.8 (Hansen Obligations). *Given a Hansen model $\Delta^H = (\mathcal{W}, I, L, <)$, we say that $O(A)$ follows if, and only if, for every total order $\prec$ extending $<$, for the Hansen-Brewka tuple (I', L') , $\text{Conclusion}(I') \cup \mathcal{W} \vdash A$.⁵¹*

Despite the similarities, there are significant differences between Hansen’s approach and ours; both in the formalism and in the field of application. The most notable, of course, is that Hansen does not have any distinction between requirements and oughts as he is only interested in what effect permissions, or licences, have on what one is required to do. Oddly enough, as these requirements arise purely out of the set of imperatives in Hansen’s model, they are formally more closely related to our oughts rather our requirements, even though, in terms of their intended interpretation, we believe, they are more closely related to requirements. But on the formal side, we also find two important differences that we have already mentioned and will discuss in more detail now.

First, Hansen’s use the “von Wright consistency” for the tuples of Imperatives and licences leads to different results. Second, even if one ignores this difference, as there is no analogue of our stable scenarios, the results of Hansen’s and our model still diverge significantly. Let’s start with the first part. We will show that Hansen’s model generates different results for both our oughts and our requirements.

We first show it for requirements. First, consider the Hansen model $(\mathcal{W}, R_{H1}, J_{H1}, <)$ that we interpret as a (dual-role) context as $\Delta_{H1} = (\mathcal{W}, R_{H1}, R_{H1} \cup J_{H1}, <)$. Let $R_{H1} = \{!c\}$ and $J_{H1} = \{!a, !b\}$ and let $\mathcal{W} = \{a \wedge b \rightarrow \neg c\}$. Let the partial order $<$ be empty.

It is easy to verify, that in Hansen’s model, we only obtain a single Hansen-Brewka tuple, namely (R_{H1}, J_{H1}) as both a and b are individually von Wright consistent with c and $\mathcal{W}$. That means for Hansen we get $O(c)$. On the other hand, for our model, we

⁵⁰Strictly speaking, $\text{Max}_{\prec}(\text{Active}_{\Delta^H}((I'_i, L'_i)))$ is a singleton set, and what we add here is its singleton element, but in order to keep the definition readable, we have omitted this technicality in the main text of the definition.

⁵¹Two things are notable here. First, Hansen also defines an operator for the “conflict account” for this semantics, that is, an operator where it is enough, if A can be derived from *some* such Hansen-Brewka tuple. We omit this operator here, as our account uses the disjunctive account and so comparing the two disjunctive semantics is more fruitful. Second, Hansen conditionalizes the operator on $\mathcal{W}$, that is, the operator is dyadic and looks like $O(\cdot/W)$. However, this is merely a difference in presentation and, to keep the operator close to the one we introduced, we choose to keep it monadic.

2. Supererogation and Dual-Role Views of Reasons

obtain 4 Brewka scenarios, namely $\{!c\}$, $\{!c, !a\}$, $\{!c, !b\}$, $\{!a, !b\}$. Since c does not follow from (the conclusions of) $\{!a, !b\}$, we do not get that $\Delta_{H1} \vdash \text{Must}(c)$. So Hansen and our model do not coincide for requirements.

What about for oughts? It is even easier to show a counterexample here. Consider the Hansen model $(\mathcal{W}, R_{H2}, J_{H2}, <)$ with our corresponding context $\Delta_{H2} = (\mathcal{W}, R_{H2}, R_{H2} \cup J_{H2}, <)$. Let $R_{H2} = \{!d, !e\}$, $J_{H2} = \{!f\}$, with $\mathcal{W} = \{d \wedge e \rightarrow \neg f\}$ and the partial order $<$ empty. Here, there are three Hansen-Brewka tuples in Hansen's case: $(\{!d\}, \{!f\})$, $(\{!e\}, \{!f\})$, and $(\{!d, !e\}, \emptyset)$. We can see that $O(d \wedge e)$ does not follow, as it does not follow from every set. For our model, however, as we generate oughts exclusively from (the (stable scenario containing only) the set of requiring reasons, we immediately obtain $Ought(d \wedge e)$.

We have seen that the difference between von Wright consistency and ordinary consistency makes a difference. But this is not the only difference here. We believe that the reason Hansen makes use of von Wright consistency in the first place is to get a similar result as we get from the construction with stable scenarios: Which permissions (or justifying reasons) one follows has to be up to the agent somehow. So next, we briefly show that even if we make use of ordinary consistency (and thus, use the definition of active imperatives and licences mirrors our definition of active reasons), the lack of stable scenarios still makes for different results. For the case of ought, we can just re-use our previous example, for the requirements, we do the following.

Consider the Hansen model $(\mathcal{W}, R_{H3}, I_{H3}, <)$ and let $\Delta_{H3} = (\mathcal{W}, R_{H3}, R_{H3} \cup J_{H3}, <)$ be the corresponding dual-role context. Let $R_{H3} = \{!c\}$, $J_{H3} = \{!a, !b\}$, and $\mathcal{W} = \{a \wedge b \rightarrow \neg c\}$. This time, we set $<$ as the smallest partial order such that $a > c$ and $b > c$. Now if we apply our original definition of active reasons and apply it to the Hansen model, we obtain a single Hansen-Brewka tuple for Hansen, namely $(\emptyset, \{a, b\})$. In our case, on the other hand, we obtain four Brewka scenarios: $\{!c\}$, $\{!c, !a\}$, $\{!c, !b\}$, $\{!a, !b\}$. This means that for our model, we can derive $\text{Must}(\neg(a \wedge b) \rightarrow c)$, but as $\neg(a \wedge b) \rightarrow c$ does not follow from $\emptyset \cup \mathcal{W}$, we do not get $O(\neg(a \wedge b) \rightarrow c)$ in Hansen's case.

So we have seen that while Hansen's model of imperatives and licences is structurally quite similar to ours, it does not produce the same results even for simple cases. This is because Hansen's account has two differences, each of which individually suffices to account for some differences. First is Hansen's use of von Wright consistency instead of ordinary consistency, second is the lack of something corresponding to our stable scenarios. Whether Hansen's approach is a good way to model explicit permissions is not something we will take a stand on here. However, we do believe that our model can be used to study permissions, and this is something we aim to do in future work. As a model of reasons, supererogation, and oughts and requirements, however, we believe Hansen's approach is insufficient.

In this section, we have compared our approach to the formal models of Mullins, McN-

mara, and Hansen. We have argued that Mullins' approach fails to resolve the puzzling cases and associated paradoxes and should thus be rejected. In McNamara's case, we have argued that by being able to account for the puzzles, McNamara's logic loses plausibility and explanatory power, whereas our model does not. We also argued that despite some formal similarity, Hansen's model cannot accommodate requiring and justifying reasons as its non-standard consistency notion and lack of an analogue of our stable scenarios makes it unsuitable for this task. In the next section, we will look at things our model cannot do well.

2.7. Shortcomings

In this section, then, we will look at weak points of our model. Specifically, these concern the impossibility of aggregating reasons, where, for example, two weaker reasons can jointly defeat one stronger reason, and the limitations in how the context can change the strength of reasons.

2.7.1. Aggregation

Consider the following scenario:

You are invited to a party. You absolutely hate the kind of music that will be played there, but on the other hand, a friend you haven't seen in a while will be there and the food will be free and pretty good. You can either go to the party or not.

We are confronted with three reasons, two speaking in favour of going to the party – that your friend will be there and that there will be good free food – and one against – that the music will be horrible; For this example, we ignore any distinction between justifying and requiring reasons. In this case, we believe that each of the reasons for going to the party alone would not be enough to make it the case that you ought to go; the music is just too unbearable. But taken together, they make it the case that you *do* ought to go. This phenomenon, where reasons *aggregate* or *accrue*, is uncontroversial as far as it goes, but our model and other similar models like (Horty, 2012) and (Mullins, 2021) cannot properly account for it.⁵² On Horty's preliminary account, a stronger reason will defeat all incompatible weaker reasons, while in our case, a stronger reason prevents the inclusion of the weaker reasons in Brewka scenarios. In either case, one strong reason is enough to exclude all incompatible weaker reasons from consideration.

⁵²In fact, our model, as stated cannot even have two separate reasons for the same action, but extending it to allow for that is trivial.

2. Supererogation and Dual-Role Views of Reasons

Of course, one can remedy this by way of modeling. For example, if one includes not only the two individual reasons for going to the party, but another reason, say “the fact that my friend will be there and there is good free food”, that is both stronger than the two individual reasons and the reason against going to the party, one can get the desired results, but this has the air of cheating. Ideally, in a logic of reasons, one should specify only the *basic* reasons and how reasons ultimately aggregate to determine oughts and requirements should be handled internally.

2.7.2. Holism

In fact, the problem of aggregation is only one part of a larger incongruity such models and the way reasons are commonly talked about in the non-formal philosophical literature. Here, the most often employed non-formal model (or, perhaps, analogy) is the kitchen scale model: reasons come with weights, they accrue much like physical weights add up. And, in the simplest version, you ought to perform an action if, and only if, the joint weight of the reasons for that action is larger than the joint weight of the reasons against this action. A formal exploration of this model will be the topic of the next chapter. If one independently thinks that formal models in the style of the ones we presented are a competitor for the kitchen scale model – (Drai, 2018) argues that they are – they leave something to be desired that the kitchen scale model can account for easily. Sometimes, it seems the context a reason appears in can change its strength (or weight, if you follow the kitchen scale model).⁵³ Horty proposes a way to include this explicitly in the logic. If we translate his idea to our model, we would include formulas of the form $!(Y < !X)$ which is a normative statement saying that reason $!Y$ is less important than reason $!X$. That is, Horty lifts the partial order from the semantics to the language. And while this has the appearance of doing what we want, there are subtle differences. For one, statements of this form are always *comparative*. They provide us with information about the relative weight of two reasons. The way context shifts the strength of reasons, however, seems more atomic. The strength of a certain reason increases or decreases in some context independent of all other reasons. Of course, this can be represented by simply stipulating that the reason now is ranked higher than all other reasons (or all other reasons up to a certain strength). But this remains vague, and a better way of modeling context shifts is needed. So, in the next chapter, we will address just this challenge.

⁵³This position is usually called *holism*. And just how much the strength of reasons can vary is subject to some debate. We will explore holism in more detail in the next Chapter.

2.8. Conclusion

In this section, we will briefly summarize the chapter and point at what we consider fruitful avenues for future research.

So far, we have presented a simple model of reasons. We have shown that this model struggles to faithfully represent cases of supererogation. We have extended the model to include a distinction between requiring and justifying reasons. We then presented three paradoxes surrounding the concept of supererogation. We have shown that our model can resolve each of the paradoxes, but we have also seen that one of them, the Intransitivity Paradox, can be tackled without needing to distinguish the two types of reasons. We then compared our model to existing approaches in the philosophical literature and argued that our model lends support to the thesis that requiring reasons are always also justifying reasons. Next, we have compared the model to two recent accounts of supererogation and argued that both of them are ultimately inferior to the model we propose. We also highlighted similarities and differences to work on permission in deontic logic. Lastly, we briefly discussed two shortcomings of the model: the fact that it cannot naturally account for the aggregation of reasons and that it struggles to represent the way context can affect the strength of reasons.

We think the way the model handles the puzzles and paradoxes surrounding supererogation should make us optimistic that it can handle other philosophical puzzles as well. Specifically, we think that it can help shed light on the various problems surrounding permissions, such as distinguishing when a permission is a *free choice* permission and when it is not (see (Kamp, 2013)). This claim is bolstered by the fact that the way permission has been modeled (notably by (Hansen, 2014), who we discussed above) is technically quite similar to the model we have presented. We conjecture that what makes a permission a free choice permission is that it is essentially grounded in exclusively justifying reasons. To defend this view, and to see if it is as plausible as we think, however, a lot more work, and modifications to the models, are needed.

Additionally, we think that the models of reasons introduced in Section 1.5 could benefit from being extended so they can represent dual-role views of reasons. (Tucker, forthcomingb) argues that this might be easy for quasi-decision-theoretic models like (Sher, 2019) (and, we would argue, (Dietrich and List, 2011)), and we tend to agree. For models based on justification logic, however, a lot more work is needed to allow these models to naturally incorporate requirements and oughts in a way that is as intuitive as the one we present here.

One avenue of future research, however, will not have to wait long to be explored: How can we model reasons in a way that avoids the shortcomings we discussed in Section 2.7? In the next chapter, we will propose a model of reasons that allows us to stay close to the *kitchen scale model*. Reasons come with a weight that can be added up and that

2. Supererogation and Dual-Role Views of Reasons

can change depending on the context. We will see that in this model, too, we can distinguish between justifying and requiring reasons in a very natural way. This model and its applications will be the topic of Chapter 3.

3. The Kitchen Scale Model Formalized

3.1. Introduction

We concluded the previous chapter with an observation. Reasons have weights, these weights can be *added up*. But it is not as simple as adding up numbers. The weights of reasons (and, in particular, their combined weight) can change depending on the context they are found in. What determines if we are permitted, forbidden, or required to do an action in some situation is then determined by the joint weight of the reasons for and against the action and its alternatives. This is often metaphorically described as the “kitchen scale” model, which we explain in more detail below.¹

In this chapter, we build a formal model that explores these issues. We then use it to analyze some philosophical puzzles surrounding the weight of reasons. What does it mean for a reason to have a particular weight? Can we learn what these weights are? Are the weights of reasons really a necessary element of normative inquiry?

So, the research questions we address are the following: First, what does a formal model of normative reasons look like that takes both the idea of weight and the idea that the weights of reasons depend on context factors seriously while staying faithful to the (non-formal) philosophical literature? Second, can we give a satisfying explanation of what it means for a reason to have weight? And can we make plausible how we could ever come to learn it?

To answer these questions, we proceed as follows: in Section 3.2, we briefly revisit the notion of a reason, focusing particularly on the way context can affect a reason’s weight. We will argue that this is best made sense of by positing a *default weight* that a reason has independently of any context and a *final weight* that does depend on context. In Section 3.3, we set up the formal model. This section has three large parts. In the first part, we establish the *structure* of the model and give a preview of how the model works. In this part, we explore how context factors and reasons relate. In the second part, we see how we can derive the final weight of a reason based on the structure a context

¹The term “kitchen scale model” comes from (Dancy, 2004), who argues against using this metaphor. The model, as we described it, is implicitly or explicitly found in many places in the philosophical literature, but the best place to get some orientation is probably the papers collected in (Lord and Maguire, 2016).

3. The Kitchen Scale Model Formalized

provides and the default weights. In the last part, we then look at how deontic statuses are determined based on the final weights of reasons. In Section 3.5, we look at some challenges posed to any account of reasons that makes use of weights. In particular, we discuss whether how we calculate the final weights and *weigh* the reasons is arbitrary and whether how weights are given to reasons is arbitrary. We will argue that, on both counts, the answer is no, providing provable results. In Section 3.6, we discuss how the model can be extended to include the distinction between justifying and requiring reasons, like we did in Chapter 2. In Section 3.7, we compare our model to related work. Here, we compare the way we derive the final weights of reasons to the weighted argumentation model of (Amgoud et al., 2022; Amgoud and Ben-Naim, 2016; Amgoud and Ben-Naim, 2018); the way we derive deontic statuses (particularly using justifying and requiring reasons) to the (non-formal) account of (Tucker, forthcominga); and our model briefly to the decision theoretic framework of (Sher, 2019). In Section 3.8, we discuss fruitful avenues of future work and conclude. Like for the preceding chapter, all proofs can be found at the end in Appendix A.3.

3.2. Reasons and Holism

3.2.1. Revisiting Reasons

Let’s revisit what we said about reasons in Chapter 1 and at the very end of the previous chapter. A reason, at least for the purposes of this thesis, is (i) a fact (ii) that stands in a specific (dis)favouring relation to (iii) some action (of an agent). In Chapter 2, (i) and (ii) were modeled only implicitly. In this chapter, while the agent for whom a reason is a reason is still not explicitly included, both the fact that grounds a reason and the favouring relation will be modeled explicitly.

But the main focus of this chapter will be on something different. At the end of the previous chapter, we’ve given some examples where two weaker reasons in favour of an action can together defeat a stronger reason against it. We’ve alluded to the standard account found in the (non-formal) philosophical literature to explain this phenomenon: *the kitchen scale model of weighing reasons*. In this chapter, we will present a formal version of this model and defend it against criticism. Before starting with the formal account, we will briefly recap and slightly elaborate on the main ideas of the kitchen scale model.²

According to the kitchen scale model, reasons come not only with a polarity – saying whether they are a reason for or against some action – but also with a magnitude – saying how weak or strong of a reason they are. This magnitude is treated as an analogue of

²For the most complete (non-formal) treatment of the kitchen scale model, see (Tucker, forthcominga).

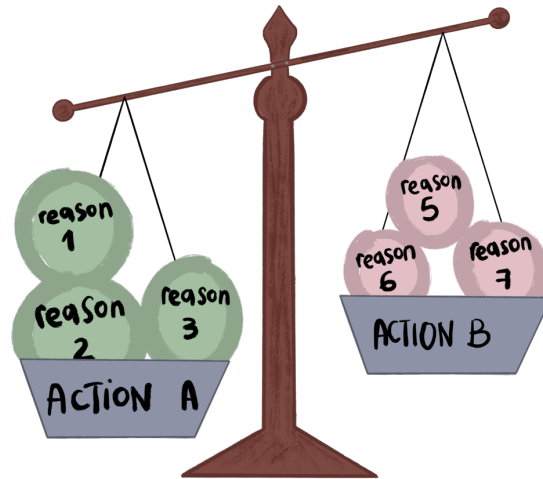


Figure 3.1.: The Kitchen Scale Model

physical weight.³ Therefore, for this chapter, we will simply call the magnitude of a reason its *weight*.⁴ The weight of two reasons, much like the weight of physical objects, can then be *added up* and this is taken to explain how and why two weaker reasons can jointly defeat or *outweigh* a stronger reason. This adding up is the phenomenon of reasons *accrual* or *aggregation*.

On the simplest account, which actions are required, permitted, or forbidden is then determined by a *weighing scale* (see Figure 3.1). In a choice between two actions, the reasons for the first action go on the left pan of the scale and the reasons for the second action go on the right pan. The joint weight of the reasons for the first action pushes the left pan down (and the right pan up) and the joint weight of the reasons for the second action pushes the right pan down (and the left pan up). Then, if the left pan is lower than the right pan, the first action is obligatory and the second action is forbidden, if the right pan is lower than the left pan, the first action is forbidden and the second action is obligatory and if both pans are balanced, both actions are permitted. With a little imagination, we can also add reasons against to a pan. Here, they add *negative weight* to it, they make the pan go up instead of down.⁵ We will see this in more detail in Section 3.4.2.

³Sometimes also couched in the nomenclature of linear algebra, where reasons are vectors comprised of a direction (polarity) and length (magnitude). (Tucker, forthcominga) gives a clean overview of why these two metaphors ultimately collapse.

⁴Sometimes, the weight of a reason is considered the combination of its magnitude and polarity. We will *not* follow this convention here. When we say weight, we mean exclusively a reason's magnitude.

⁵This model can be naturally extended to situations where more than two actions are in play. Indeed, this is how we formulate it from the start in Section 3.4.2.

3. The Kitchen Scale Model Formalized

On the face of it, this metaphor does not seem appropriate to model practical reasons. Adding physical weight is a matter of adding numbers, combining reasons is not. Reasons can be excluded from consideration, if they are outright defeated, perhaps, their weight can change from context to context⁶ and the weight of two reasons together might be weaker than either individually.⁷

Why is it, then, that anyone would take the kitchen scale metaphor as a starting point for reasons? To answer this question, we need to take a little detour and explore *Reason Holism and Atomism*.

3.2.2. Holism and Atomism

(Dancy, 2004) argues for a (in)famous thesis. What we ought to do is based on reasons, but there is no principle that explains systematically how. Something that is a reason for an action *A* in one context can be a reason against the action in another. The strength of reasons, their magnitude, also varies with contexts. Dancy's view is the combination of two distinct theses: First, *Holism*, the thesis that the weight (and potentially polarity) of a reason can vary depending on the context (the rejection of this thesis is called *Atomism*) and second, the thesis that it is impossible to find a (set of) general principles to explain how context changes the weight (or polarity) of reasons, we call this set of claims *Particularism*.⁸

We can consider this to be one end of the Holist position (let's call it *Strong Holism*). The other end is Atomism. Here, context does not change the weight (or polarity) of reasons, and presumably we can find a principle that explains how reasons combine to ground oughts. The latter seems not particularly promising, as it robs reasons of their non-monotonic properties. The former, however, is also counterintuitive. It leaves the question of how we can ever be motivated to act, know what action is the right one, figure out the weight of reasons, unanswered.⁹ Additionally, if there is no general model of how the weight of reasons changes depending on context, then certainly there is no formal model of reasons aggregation to be made. Formal ethics would be a non-starter.

We are thus lucky that the most plausible position lies in the middle. We accept that reasons can change their weight depending on context, but we deny that this is all we can say about the matter. The weight of reasons depends on the context in a very specific way that we can formally model. We call this position *Constrained Holism*. We will

⁶This is indeed what we will be arguing for in the next section.

⁷See also Section 3.5 for more challenges.

⁸These terms come from (Dancy, 2004), but we define them slightly different here. In particular, on our definition of Holism, it is not necessary that reasons can change their *polarity*, but (just) their magnitude.

⁹For a defence of Dancy's view by the man himself that still gives a good overview, see (Dancy, 2017). For a forceful attack against this kind of particularism, see (Berker, 2007).

3.3. A (Formal) Holist Model of Weighing Reasons

describe it in more detail in the next section.

3.2.3. Default Weights, Final Weights and a Constrained Holism

This constrained holism, as we understand it, comes in two parts. The first claim is that every reason comes with a *default weight* that it has independent of any context. The second claim is that it has a context-dependent *final weight* that depends on the default weight and specific context factors in a specific way.¹⁰ Only the final weight of a reason goes on the pans of the scale, determining the morally correct actions.

That leaves two questions to be answered: What are these specific context factors, and in what specific way do they determine the final weight of reasons? The first one, we will be able to answer easily as it is well-established in the literature, answering the second question will be the topic of Section 3.4.1. Context factors having an impact on the final weight of a reason can be divided into modifiers that change the weight of a reason and undercutters that remove the weight of a reason entirely, effectively making it cease to be a reason altogether. Modifiers in turn are routinely split into two types: amplifiers – context factors that *increase* the weight of a reason – and attenuators – context factors that *decrease* the weight of a reason. These four types of normative elements, reasons, undercutters, amplifiers, and attenuators, turn out to be the only elements we need to give a comprehensive model of how reasons determine what one ought to do.¹¹

In the next section, our goal is to turn the constrained holism into a proper formal model.

3.3. A (Formal) Holist Model of Weighing Reasons

In this section, we will develop a formal version of the kitchen scale model. It will contain four elements. The first element is the *structure* of normative situations. The structure models how reasons, modifiers and undercutters interact. The second element is modeling how reasons (and modifiers and undercutters) get their weight, that is, the process of *weighting* reasons. It is formally the easiest to model out of the four elements, but, as we will see in Section 3.5.2, it might well be the most contested part of the kitchen scale model in the (non-formal) philosophical literature. The third element models how

¹⁰For a detailed account and defence of such a constrained holism, see (Tucker, 2023, forthcominga). We will compare our proposal to Tucker's account in more detail in Section 3.7.

¹¹We will, in the next section, give a slightly more general picture, that allows for chained modifiers and undercutters; that is a modifier might be modified by yet another modifier which is in turn undercut by an undercutter and so forth. These concepts will become more tractable once the formal model is in place.

3. The Kitchen Scale Model Formalized

we can derive the final weights of reasons from their default weights and the structure of a normative situation. The fourth and last element, then, models how final weights determine the deontic statuses of actions. This is the business of *weighing reasons*.

3.3.1. Structure: Reasons, Considerations & Normative Graphs

In this section, we will spend a considerable amount of time setting up these four elements. We will start with the first, and most basic: the structure of normative situations. That is, we need to model reasons, modifiers, and undercutters. Once the structure is in place, we will then introduce some very simple prototypes for the other two elements, so that one can glimpse how our model, in principle, works. We will analyze (in much more detail) how else these last two elements could look like in Sections 3.4.1 and 3.4.2.

To get us started, consider the following toy scenario.

The Cake When you visited your mother yesterday, she gave you a piece of cake to take with you and made you promise that you'd eat it. Now you're on a video call with her, and the cake is in front of you. Next to you is your niece with a hungry look in her eyes. You would take delight in eating the cake, and so would your niece. You also know that your niece stole a toy from another child earlier today. Your only options are (this is a toy scenario!) to either eat the cake yourself, or to give it to your niece.

In this example, there are two possible actions and five reasons. First, the fact that you made a promise is a reason for eating the cake. Second, the fact that your niece is hungry is a reason for giving the cake to her. Next, the fact that you would take delight in eating the cake is a reason for eating the cake, while the fact that your niece would take delight in eating the cake is a reason for giving the cake to her. Finally, the fact that your niece stole a toy, and it wouldn't be fair to reward stealing with cake, is a reason *against* giving the cake to your niece.

The first normative component we will model is reasons. To model it, we need three things. First, we need to include the fact that grounds the reason, its *premise*. It is usually this fact that philosophers refer to as "the reason". Second, we need to include what action it is a reason for or against – its *conclusion*. Third, we need to include its polarity – whether it is a reason *for* or *against* an action – its *role*. We hence define the concept of a reason as follows. Note, that so far, we have not given a weight to the reason.

Let L be a non-empty (potentially uncountable) set of symbols called propositions.¹²

¹²We use the letter L here as it is often used to denote a logical language. Indeed, we could use the set of propositions of some logic here, but for our purposes it is enough to simply treat it as a set of symbols. More importantly, we will think of L as a fixed set for the entirety of this chapter that includes all the various symbols we use in examples. We also assume that it is *large enough* to find

3.3. A (Formal) Holist Model of Weighing Reasons

Definition 3.3.1 (Reasons). *We call a tuple (p, c, r^+) or (p, c, r^-) a reason if both p and c are elements of L with r^+ and r^- being special symbols representing that the tuple is a reason for or against, respectively.*

Coming back to our example, we can formalize the reasons it contains using this definition. We will use *EatCake* and *GiveCake* to represent the two actions available to you.

First, letting *Promise* stand for the fact that you promised to eat the cake. We can then model the corresponding reason as the tuple $\delta_1 = (Promise, EatCake, r^+)$. Second, letting *Hungry* stand for the fact that your niece is hungry, we can represent the corresponding reason as $\delta_2 = (Hungry, GiveCake, r^+)$. Next, letting *TakePleasureYou* stand for the fact that you would take delight in eating the cake while *TakePleasureNiece* stands for the fact that your niece would take delight in eating the cake, we can model the two corresponding reasons as, $\delta_3 = (TakePleasureYou, EatCake, r^+)$ and $\delta_4 = (TakePleasureNiece, GiveCake, r^+)$. Finally, letting *StoleAToy* stand for the fact that your niece stole a toy, we model this reason against as $\delta_5 = (StoleAToy, GiveCake, r^-)$.

We have given a formal gloss of the notion of a reason. What about amplifiers, attenuators, and undercutters? To see them in action, we will expand upon the example a little.

The Cake (extended) You know that your niece sincerely regrets having stolen the toy. What's more, as soon as your mother sees the hungry look in your niece's eyes, she releases you from the promise you had made to her.

Arguably, these two new facts do not represent reasons, as they neither speak for nor against performing an action. Instead, the fact that your niece regrets her theft acts as a *modifier*. It makes the reason given by the fact that she stole a toy less weighty. That is, it serves as an *attenuator*. The fact that your mother released you from your promise, on the other hand, serves as an *undercutter*. It completely nullifies the (weight of the) reason given by the fact that you promised. Next, we will include these modifiers and undercutters in the formal model. We will use the catch-all term *normative consideration* – or just *consideration* – to encompass reasons, modifiers and undercutters.

Definition 3.3.2 (Normative Considerations). *The set of normative considerations is the smallest set such that*

- *All reasons are normative considerations.*
- *If δ is a normative consideration, $p \in L$, and $x \in \{u, a^+, a^-\}$, then (p, δ, x) is a*

new propositions not included in any Normative Graph so far, when we need them.

3. The Kitchen Scale Model Formalized

normative consideration. Like in the case for reasons, u, a^+, a^- are special symbols representing the fact that p is an undercutter, amplifier, or attenuator of δ .

So, we can now model the amplifier and the undercutter from our extended example. The fact that your niece regrets her theft (*Regrets*) attenuates the reason $\delta_5 = (StoleAToy, GiveCake, r^-)$. We represent this as $\delta_6 = (Regrets, \delta_5, a^-)$. Which, if we spell it out fully, is the tuple $(Regrets, (StoleAToy, GiveCake, r^-), a^-)$. The fact that your mother releases you from your promise (*Release*) is an undercutter of the reason $\delta_1 = (Promise, EatCake, r^+)$, which we can now represent as $\delta_7 = (Release, \delta_1, u) = (Release, (Promise, EatCake, r^+), u)$. We represent this structure graphically in Figure 3.2.

It is nice to have a shorthand to talk about the different elements of normative considerations. The following functions handle this task.

Definition 3.3.3 (Premises, Conclusions, Roles of a Consideration). *Given a consideration $\delta = (p, c, x)$, we define:*

- $Premise(\delta) = p$
- $Conclusion(\delta) = c$ (here c is either an element of L if it is a reason or another consideration if it is not)
- $Role(\delta) = x$ (per definition $x \in \{r^+, r^-, u, a^+, a^-\}$)

We extend these functions to sets of considerations in the obvious way.

In order to model normative situations, we only need one further ingredient. In any situation, there will be a set of actions an agent can choose from. For now, we will simply assume that these choices are exhaustive and mutually exclusive. Next, we model such a normative choice situation. An agent can choose between different actions, each action might have reasons speaking for or against it, perhaps modified or undercut by other context factors (which, in turn, could be themselves modified or undercut by further considerations). We capture this network of normative dependency with a structure we call a *Normative Graph*.

Definition 3.3.4 (Normative Graph). *A Normative Graph N is a tuple (Δ, O) such that:*

- $O \subseteq L$ is a finite non-empty set called (the set of) actions, and the elements of O , representing possible actions, are called options.

3.3. A (Formal) Holist Model of Weighing Reasons

- Δ is a set of considerations.

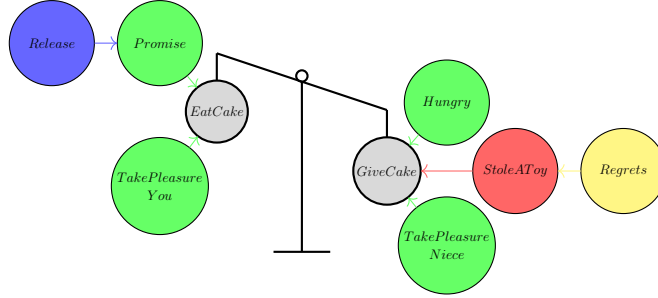


Figure 3.2.: A graphical depiction of the Normative Graph capturing the toy scenario. Options are given in grey. Considerations are labelled by their premise. Reasons for an action are given in green, reasons against in red, undercutters are modeled in blue and attenuators in yellow. The arrows point towards the conclusion of a consideration.

The reason we call such a structure a *graph* is because it can easily be mapped to an ordinary directed graph.

Definition 3.3.5 (Directed Graph of a Normative Graph). *Given a Normative Graph $N = (\Delta, O)$, its directed graph is a graph G where the nodes of G are the elements of Δ and there is a directed edge between node δ_1 and δ_2 iff $\text{Conclusion}(\delta_1) = \delta_2$.*

That leaves one element unaccounted for so far: The (default) weight of reasons, and other normative considerations. We assign weights to these normative elements via the following function.

Definition 3.3.6 (Weighting Function). *Let the set NC be the set of normative considerations as defined in Definition 3.3.2. Then we call a function $w : NC \rightarrow \mathbb{R}_{\geq 0}$ a weighting function. We will call $w(\delta)$ the (default) weight of a consideration δ (relative to w).*

Most definitions in the remainder of this section will be defined relative to some specific weighting function. Still, whenever it is immaterial which one we choose, we will usually omit mentioning it explicitly. Later in this section, we will discuss the connection between the default weights (via the weighting function) of a reason and its context-dependent final weight.

We have seen how we can model normative scenarios as a Normative Graph. But not every Normative Graph will represent a possible scenario. Some Normative Graphs as

3. The Kitchen Scale Model Formalized

we have defined them above will definitely not correspond to any scenario, while whether others do is subject to debate. In this section, then, we will make this precise. We will present two sets of constraints on graphs. The first set excludes all graphs that are intuitively wrong, while the second set characterizes some graphs that one might, or might not, want to exclude depending on one's metaethical commitments.

We will collect the set of *reasonable* graphs, the graphs we consider possible, in the set $\mathcal{N}$ and then model the constraints as constraints on this set $\mathcal{N}$. We will call these constraints *principles* for reasons that will become apparent in the next section of this chapter.

For the constraints, we need to introduce some further definitions. First, we want the graph to be *complete*, that is, we do not want it to contain any modifiers or undercutters of considerations that are not themselves part of the graph. The same goes for reasons; every reason in the graph should be a reason for an action that can be performed in the situation.

Definition 3.3.7 (Complete Normative Graph). *We call a Normative Graph $N = (\Delta, O)$ complete iff for every $\delta \in \Delta$, $\text{Conclusion}(\delta) \in O \cup \Delta$*

This gives us the principle:

Principle 3.3.1 (Completeness). *Every Normative Graph in $\mathcal{N}$ is complete.*

Next, as briefly discussed in Chapter 2, we expect any Normative Graph to only include a finite number of reasons (and other normative considerations). We can express this as a principle on $\mathcal{N}$:¹³

Principle 3.3.2 (Finite). *The directed graph of every Normative Graph in $\mathcal{N}$ is finite.*

Lastly, we do not allow for any cycles to be formed in the Normative Graph, as practical normativity simply doesn't seem to allow for them.

Principle 3.3.3 (Acyclicity). *The directed graph of every Normative Graph in $\mathcal{N}$ is acyclic.*¹⁴

¹³Should there be cases with an infinitude of actions or considerations available to the agent, then surely these are exceptional cases. In this case, we will only model the normal kind of case and leave cases with an infinite number of actions or considerations for future work.

¹⁴We might want a stronger notion of acyclicity here, where we take premises into consideration as well.

3.3. A (Formal) Holist Model of Weighing Reasons

The principles we have introduced so far, are necessary, we claim, to model any normative situation at all. The principles in the next set are subject to debate. We want to remain agnostic about whether to accept them or not, and so, for the remainder of this thesis, we will not restrict the set of Normative Graphs any further.

First note, that we can model the atomist position simply by constraining the set of Normative Graphs we deem acceptable.¹⁵

Principle 3.3.4 (Atomism). *Every graph in $\mathcal{N}$ contains neither modifiers nor undercutters.*

Of course, we can also deny atomism by a position we will simply call holism here.¹⁶

Principle 3.3.5 (Holism). *Some graphs in $\mathcal{N}$ contain modifiers or undercutters.*

We can also distinguish what kind of (constrained) holism we allow. One such distinction that will become important in Section 3.5.1 is whether we allow that a fact that grounds a reason can also ground a modifier or undercutter. To illustrate this fact, imagine you are to decide between going to work or faking being sick while staying at home. You have (this a contrived example) a single reason to stay home, namely that you have a lot of work to do (and it would be unpleasant to do it). You also have a single reason against staying at home, namely that your colleagues would have to do your work for you. It is, at least somewhat, plausible that the fact that you have a lot of work actually amplifies this reason. After all, the more work you have, the more work your colleagues would have to do. If we allow for these kinds of cases, where the same fact grounds both a reason and an amplifier, we need *inclusive* Normative Graphs.

Definition 3.3.8 (Inclusive Graph). *A (holist) graph is called inclusive iff there is a reason δ_1 and a normative consideration that is not a reason δ_2 in the graph such that*

We will go through two such scenarios. First, consider the case of two reasons, $r_1 = (p, q, r^+)$ and $r_2 = (q, p, r^+)$. This structure is allowed according to our definition so far. We can read this as p is a reason for q (-ing) and q is a reason for p (-ing). If we wanted to exclude these kinds of cyclical structure, we could posit that in a Normative Graph, no element of the actions O can figure as the premise of a consideration. Second, consider the following case where we have two considerations, δ_1 and δ_2 , where $\delta_1 = (p, \delta_2, a^-)$ and $\delta_2 = (p', p, r^+)$. This is a kind of cycle too, where p is a modifier of p' , which in turn is a reason for p . There might be good reasons to want to exclude these cycles too. However, we will not introduce these as further constraints, as we are not aware of any precedent in the philosophical literature that warrants their exclusion or inclusion.

¹⁵Note that, strictly speaking, this only models the atomist position if one additionally accepts that the only way the context-dependent weight of a reason can differ from its default weight is mediated through undercutters and modifiers. That is, if one accepts what we have called a *constrained holism*.

¹⁶Note, though, that as explained in Section 3.2.2, holism comes in many shades and what we will suggest here is what we called a *constrained holism*.

3. The Kitchen Scale Model Formalized

$$Premise(\delta_1) = Premise(\delta_2).^{17}$$

We can then model the two positions one can take on the matter.¹⁸

Principle 3.3.6 (Exclusive Holism). *There are no inclusive graphs in $\mathcal{N}$.*

Principle 3.3.7 (Inclusive Holism). *There is at least one inclusive graph in $\mathcal{N}$.*

Lastly, we might take the position that while modifiers and undercutters can modify or undercut reasons, they themselves cannot be undercut or modified.

This leads us to the next principle.

Principle 3.3.8 (Single-Level Graph). *For every Normative Graph N in $\mathcal{N}$, N contains no modified or undercut modifiers and no modified or undercut undercutters.*

Alternatively, we might take the more liberal position and say that modifiers and undercutters can themselves be modified or undercut by modifiers or undercutters, which in turn can be modified or undercut and so forth.

Principle 3.3.9 (Multi-Level Graph). *There is at least one Normative Graph N in $\mathcal{N}$, that contains a modified undercutter or modifier, or an undercut modifier or undercutter.*

With these principles in place, we have completed the *structural* part of our model. What remains to be done is to (i) determine the final weights of reasons and (ii) determine the deontic statuses of actions based on those final weights. We will discuss both of these in turn.

3.3.2. Final Weights & Weighing Reasons

Our first goal is to introduce two functions. First, a function that delivers the final weight of considerations based on the structure of a graph. The topic of the next section(s) will be discussing criteria of adequacy for such a function. For now, however, to illustrate the model, we will use the simplest function that might do the job. We call this function the

¹⁷We borrow the term inclusive and exclusive holism from (Tucker, forthcominga).

¹⁸The reason inclusive holism is favoured by some philosophers is that it allows to model the phenomenon of diminishing marginal utility (see (Tucker, 2023, forthcominga) and Section 3.5.1.

3.3. A (Formal) Holist Model of Weighing Reasons

Additive Function or f_1 . The basic idea behind this function is that an amplifier simply adds its strength to the consideration it amplifies, and an attenuator simply subtracts its strength. An undercutter, as long as it has positive weight, will reduce the weight of a consideration to zero. To make this formal, we will introduce some shorthand notation.

First, we need to reference the set of amplifiers, attenuators and undercutters of some consideration.

Definition 3.3.9 (Modifiers and Undercutters of a Consideration). *Given a Normative Graph $N = (\Delta, O)$ and a consideration $\delta \in \Delta$, we define*

- $A_\delta^+(N) = \{\delta^* \in \Delta \mid \text{Conclusion}(\delta^*) = \delta \text{ and Role}(\delta^*) = a^+\}$
- $A_\delta^-(N) = \{\delta^* \in \Delta \mid \text{Conclusion}(\delta^*) = \delta \text{ and Role}(\delta^*) = a^-\}$
- $A_\delta^u(N) = \{\delta^* \in \Delta \mid \text{Conclusion}(\delta^*) = \delta \text{ and Role}(\delta^*) = u\}$

Where the context makes clear what Normative Graph we are talking about, we omit the Normative Graph, writing, for example, just A_δ^+ for the set of amplifiers of some consideration δ .

Next, we introduce some helpful shorthand to make the definition of our function more concise.

Definition 3.3.10 (Undercutter Shorthand). *Let δ be some consideration in some Normative Graph N , A_δ^u be the set of undercutters of δ , then we define:*

$$u(\delta) = \begin{cases} 1 & \text{if } \sum_{\delta' \in A_\delta^u} f_1(\delta') = 0 \\ 0 & \text{otherwise} \end{cases}$$

With this shorthand, we can define the Additive Function.

Definition 3.3.11 (Additive Function).

$$f_1(\delta) = \max(0, w(\delta) + \sum_{a^+ \in A_\delta^+} f_1(a^+) - \sum_{a^- \in A_\delta^-} f_1(a^-)) * u(\delta)$$

As our Normative Graphs are acyclic, both the shorthand and the function are well-defined. We add the condition that the final weight of a consideration cannot drop

3. The Kitchen Scale Model Formalized

below zero, to keep the intuitions about physical weights alive. We discuss this property in more detail in Section 3.4.1.

Having a simple function to determine the final weights of a consideration in place, we next need to determine the deontic statuses of actions. Here too, we propose a simple function that we think corresponds rather well to the kitchen scale metaphor. Given a Normative Graph $N = (\Delta, O)$, this function returns a subset of O , the set of permissible actions given this graph. If an action is not in this subset, we call it forbidden, and if there is a single action in the subset, we say that this action is required. The function, which we call *Sum of Reasons*, simply adds up the weights of all reasons for an option and subtracts the weights of all reasons against this option. If this joint weight of the reasons for and against an option is larger than for all alternatives, this option is included in the set of permissible actions. Again, for now, we take no stance on whether this function is plausible. We analyze different functions to determine the deontic statuses of actions in Section 3.4.2.

Definition 3.3.12 (Sum of Reasons). *Let R_N be the set of reasons included in N . Then:*

$$\begin{aligned}
 g_1(R_N) = \{o \in O \mid & \text{for all } o^* \in O : \\
 & \sum_{\{r \in \Delta : \text{Role}(r)=r^+ \text{ and Conclusion}(r)=o\}} f_1(r) - \sum_{\{r \in \Delta : \text{Role}(r)=r^- \text{ and Conclusion}(r)=o\}} f_1(r) \\
 & \geq \\
 & \sum_{\{r \in \Delta : \text{Role}(r)=r^+ \text{ and Conclusion}(r)=o^*\}} f_1(r) - \sum_{\{r \in \Delta : \text{Role}(r)=r^- \text{ and Conclusion}(r)=o^*\}} f_1(r)\}
 \end{aligned}$$

These two simple functions allow us to demonstrate how the model (in principle) works. As an example, consider again the scenario introduced in Section 3.3.1, where you have to decide between eating a cake and giving it to your niece. We had represented this scenario graphically in Figure 3.2 on page 65. In order to illustrate these definitions, we use a weighting function w that, we think, roughly captures the intuitions in this case. The Normative Graph corresponding to the scenario (including weights) is given as N_1 , where:

- $\delta_1 = (\text{Promise}, \text{EatCake}, r^+)$ with $w(\delta_1) = 3$
- $\delta_2 = (\text{Hungry}, \text{GiveCake}, r^+)$ with $w(\delta_2) = 4$
- $\delta_3 = (\text{TakePleasureYou}, \text{EatCake}, r^+)$ with $w(\delta_3) = 2$
- $\delta_4 = (\text{TakePleasureNiece}, \text{GiveCake}, r^+)$ with $w(\delta_4) = 2$

3.4. A Principle-Based Analysis of Reasons and Their Final Weight

- $\delta_5 = (\text{StoleAToy}, \text{GiveCake}, r^-)$ with $w(\delta_5) = 3$
- $\delta_6 = (\text{Regrets}, \delta_5, a^-)$ with $w(\delta_6) = 2$
- $\delta_7 = (\text{Release}, \delta_1, u)$ with $w(\delta_7) = 4$

In order to determine the deontic statuses of the actions in this scenario, we need to first calculate the final weights of the considerations in the Normative Graph.

Here, we start at the leaves of the graph; their final weight is just their default weight. This means that, as $\delta_2, \delta_3, \delta_4, \delta_6, \delta_7$ are all leaves in the graph, $f_1(\delta_2) = 4, f_1(\delta_3) = 2, f_1(\delta_4) = 2, f_1(\delta_6) = 2$, and $f_1(\delta_7) = 4$.

The weight of δ_5 depends on (the final weight of) δ_6 . Since for f_1 attenuators simply subtract their final weight, we are left with $f_1(\delta_5) = w(\delta_5) - f_1(\delta_6) = 3 - 2 = 1$. For the weight of δ_1 , we need to do no calculation, as it is undercut by an undercutter of positive final weight and so $f_1(\delta_1) = 0$.

Having calculated the final weights of each consideration, this leaves us to determine the set of deontic statuses. For g_1 , we need to check (for each action) whether the sum of the reasons for this action minus the sum of the reasons against this action is larger than that of any alternatives. If it is, it counts as permitted. In our case, we have two reasons in favour of eating the cake δ_1 and δ_3 and no reasons against eating the cake. This means that we have to compare $f_1(\delta_1) + f_1(\delta_3) = 0 + 2 = 2$ to the sum for giving the cake to your niece. Here we have two reasons in favour of giving the cake to your niece and one against. This means that the sum total of the reasons is $f_1(\delta_2) + f_1(\delta_4) - f_1(\delta_5) = 4 + 2 - 1 = 5$. Since $2 < 5$, this means that $\text{EatCake} \notin g_1(N_1)$, but $\text{GiveCake} \in g_1(N_1)$. And so, in this case it is not permitted to eat the cake yourself, but it is permitted to give the cake to your niece. In fact, since giving the cake to your niece is the uniquely permitted option, it is also required.

3.4. A Principle-Based Analysis of Reasons and Their Final Weight

We concluded the last section by introducing two functions. The first, f_1 , calculated the final weight considerations based on the graph structure they are found in. The second function, g_1 , then gave us the deontic statuses of actions based on this final weight.

We have said that these are simple functions that we used because they are *simple*. In this section, we will investigate what functions we could use in their place. We will tackle the task of finding a function f that calculates the final weights of reasons and

3. The Kitchen Scale Model Formalized

other considerations first. Unfortunately, the philosophical literature is largely silent on how modifiers contribute to the final weight of reasons, and so this part of the thesis is largely exploratory. In order to explore this question in a structured way, we will perform a *principle-based* analysis of the possible ways to assign weights.¹⁹

The idea of principle-based analyzes stands in a long and fruitful tradition. Before we get going, we are going to explain the approach while locating it in its tradition. We distinguish between four different stages of principle-based analysis: (i) determining the function domain; (ii) categorizing the possible sets of functions using principles; (iii) proving statements (such as impossibility results) directly from the principles; (iv) characterizing or *axiomatizing* (sets of) functions entirely by principles. The first two stages are the base for the third and fourth, which are largely independent of one another. We will explain each stage by setting them in a historical context.

The most important early success in principle-based analyzes is probably (Arrow, 1950). Arrow famously proved an impossibility result about ranked-choice voting that says that, very roughly, no ranked-choice voting function exists that satisfies all of a set of desired principles.²⁰

The first stage consists in simply describing the domain we are interested in. In Arrow's case, this means ranked choice voting functions. In our case, it means ways to calculate the final weights of normative considerations.

The second stage then consists in laying out principles that can describe different functions in that domain, and group them according to which of the principles they satisfy and which they violate. For example, in Arrow's case, one such principle is the *Independence of Irrelevant Alternatives*, which says that, roughly, adding irrelevant voting choices has no impact on the outcome of an election. In our case, one principle we will propose is that if a reason is not modified or undercut by anything, then its final weight is simply its default weight (see Principle 3.4.2). Not every possible ranked-choice voting function satisfies the principle of *Independence of Irrelevant Alternatives* just like not every function to calculate the final weights of reasons satisfies the principle that the final weight of unmodified and not undercut reasons is their default weight.

The third stage then comprises results purely in the language of principles. Here is where we find Arrow's famous result. We, too, will prove a (much less general) result below that shows that no function exists that satisfies a certain combination of principles. The fourth stage comprises results that *axiomatize* certain functions. That is, results that show that, given a set of principles, only a single function (or a well-described set of functions) exists that satisfies every principle in the set. The prime example here is

¹⁹This will turn out to be very similar to the principle-based analysis by (Amgoud and Ben-Naim, 2016; Amgoud and Ben-Naim, 2018; Amgoud et al., 2022) who work in a similar setting. We will highlight the differences and similarities in Section 3.7.

²⁰Specifically, it states that there is no such voting function that satisfies the principles of pareto efficiency, non-dictatorship and independence of irrelevant alternatives. These results are due to (Wilson, 1972) who improved slightly on Arrow's original results.

3.4. A Principle-Based Analysis of Reasons and Their Final Weight

found in decision theory, where one can show that accepting a certain set of principles of rationality forces one to accept that the only way to aggregate preferences (described as utility functions) of different agents is a weighted sum.²¹ In our case, this would mean showing that there is only a single *specific* function or clearly delineated set of functions that satisfies a certain set of principles.

More closely related to the topic of this thesis, are principle-based analyzes in the context of formal argumentation, first done by (Baroni and Giacomin, 2007). The closest match to our analysis is due to (Amgoud and Ben-Naim, 2016; Amgoud and Ben-Naim, 2018; Amgoud et al., 2022) who work with weighted argumentation frameworks. In fact, many principles we propose can be translated to their system too and vice versa.

The principle-based analysis we perform will mostly remain in the first and second stage. We will prove one impossibility result (the proof can be found in Appendix A.3). However, this proof is mostly a proof-of-concept. We will not attempt to characterize the functions completely by axioms, and so stage four is left for future work.

3.4.1. Determining Final Weights

Principles

In this part of this section, we introduce various principles about the way final weights are determined. In the next section, we will introduce four different functions that are plausible candidates and compare them based on the set of principles. Some of the principles, like in Section 3.3.1, will be criteria of adequacy such that no function not satisfying the principles can be deemed adequate. Other principles we are more careful about. We think they carve out important differences in what kind of functions are possible, but we want to remain agnostic whether a function needs to satisfy them. Before getting to the principles, we need to introduce some further notation.

First, we will often need to discuss the graph that is generated by adding or removing a single consideration.²²

Definition 3.4.1 (Adding a Consideration). *Given a Normative Graph $N = (\Delta, O)$ and a consideration δ , we write $N + \delta = (\Delta \cup \{\delta\}, O)$.*

²¹See (Harsanyi, 1955). Specifically, these principles are monotonicity, symmetry, continuity, and locality.

²²There is an additional complication which we will ignore here, but which is important for applications. Often, when we consider cases where a reason does not hold, we do not mean to remove a tuple like we have defined it here, but the *fact that grounds* the reason. Arguably, in such a case, we would want to remove any consideration that is grounded by this same fact p and so, in this case, we would remove all considerations δ with $Premise(\delta) = p$.

3. The Kitchen Scale Model Formalized

Definition 3.4.2 (Removing a Consideration). *Given a Normative Graph $N = (\Delta, O)$ and a consideration δ , we write $N - \delta = (\Delta \setminus \{\delta\}, O)$.²³*

All of these definitions can be extended to sets of considerations in the obvious way.

Additionally, we not only want to add or remove considerations but options as well.

Definition 3.4.3 (Removing an Option). *Given a Normative Graph $N = (\Delta, O)$ and an option $o \in O$, we say that we remove o from N , symbolically $N - o$ with $N - o = (\Delta, O \setminus \{o\})$.*

Definition 3.4.4 (Adding an Option). *Given a Normative Graph $N = (\Delta, O)$ and a proposition o such that for no $\delta \in R$, $Premise(\delta) = o$ we write $N + o = (\Delta, O \cup \{o\})$.*

Here again, we can extend the definitions to sets of options.

With these definitions in place, we can now introduce the principles. But as we have not yet determined how the final weight of a consideration is generated, all we know is that it depends on the Normative Graph the consideration is located in. To be able to talk about the final weight (without explicitly giving a way to compute it), for a given graph $N = (\Delta, O)$ and a weighting function w we introduce the function $f_N^w : \Delta \rightarrow \mathbb{R}$. Where the context makes it clear what Normative Graph we are talking about, we omit the subscript on f and where the choice of the weighting function is not important, we will omit the superscript as well. Usually, then, we will simply write $f(\delta)$ for the final weight of δ .

Next, we will introduce the principles one by one. We will first describe the principle in natural language and afterwards give a formal definition.

The first principle, Anonymity, states that the *names* of actions and considerations do not play a role in determining final weights; only the structure and weights do the work.

Principle 3.4.1 (Anonymity). *Let two Normative Graphs $N = (\Delta, O)$ and $N^* = (\Delta^*, O^*)$ be called structurally identical iff there is a bijection ϕ from $\Delta \cup O$ to $\Delta^* \cup O^*$ such that $\phi[O] = O^*$ and for every $\delta \in \Delta$, $\delta' \in \Delta \cup O$, $w(\delta) = w(\phi(\delta))$ and $Conclusion(\delta) = \delta'$ iff $Conclusion(\phi(\delta)) = \phi(\delta')$ and $Role(\delta) = Role(\phi(\delta))$.*

²³Note, that it is not guaranteed that $N + \delta$ or $N - \delta$ is a *reasonable* Normative Graph according to our definition, as it is not guaranteed to be complete. We could remedy this by removing any considerations that are not or no longer on a path to an action in the corresponding directed graph. As this makes the notation quite cumbersome, however, we skip over this detail here.

3.4. A Principle-Based Analysis of Reasons and Their Final Weight

Then, for all structurally identical graphs $N = (\Delta, O), N' = (\Delta', O')$: for $\delta \in \Delta$, $f_N^w(\delta) = f_{N'}^w(\phi(\delta))$.

The next principle, Unmodified Default, was already briefly mentioned when introducing the idea of a principle-based analysis. It states that the final weight of unmodified and not undercut normative considerations is just their default weight. This is a quite weak principle, and we will see stronger versions of it with the Principles Weak Modifier Reciprocity and Strong Modifier Reciprocity.

Principle 3.4.2 (Unmodified Default). *For any weighting function w and Normative Graph N , if $A_\delta^+(N) = A_\delta^-(N) = A_\delta^u(N) = \emptyset$, then $f_N^w(\delta) = w(\delta)$.*

The third principle, Modeler's Delight, says that the final weight (like the default weight) is non-negative and that if a consideration is undercut, its final weight is exactly zero.

Principle 3.4.3 (Modeler's Delight). *For all weighting functions w and all considerations δ in some Normative Graph N :*

1. $f_N^w(\delta) \geq 0$,
2. If there is an undercutter δ^u of δ with $f_N^w(\delta^u) > 0$, then $f_N^w(\delta) = 0$.

The principle No Effects from Spurious Modifiers says that, given a Normative Graph, adding an arbitrary number of considerations of final weight 0, does not change the final weight of other considerations.

Principle 3.4.4 (No Effects from Spurious Modifiers). *Given any weighting function w , any Normative Graph $N = (\Delta, O)$, and a set of considerations X , such that for all $\delta' \in X$, $f_{N+X}^w(\delta') = 0$, then for every $\delta \in \Delta$, $f_N^w(\delta) = f_{N+X}^w(\delta)$.*

Principle No Valence Flips for Modifiers tells us that no amplifier (no matter the context it appears in) ever attenuates a consideration, and no attenuator ever amplifies.

Principle 3.4.5 (No Valence Flips for Modifiers). *For any weighting function w and any Normative Graph $N = (\Delta, O)$. Given an amplifier δ^+ of some consideration $\delta \in \Delta$, $f_N^w(\delta) \geq f_{N-\delta^+}^w(\delta)$. Analogously, for an attenuator δ^- of some consideration δ , $f_N^w(\delta) \leq f_{N-\delta^-}^w(\delta)$.*

The next principle, Immediate Successors, states that the final weight of a consideration depends solely on its default weight and the final weight of its undercutters, attenuators and amplifiers.

3. The Kitchen Scale Model Formalized

Principle 3.4.6 (Immediate Successors). *Let w be an arbitrary weighting function and $N = (\Delta, O)$ and $N' = (\Delta', O')$ be two Normative Graphs with $\delta \in \Delta \cap \Delta'$. If the sets of attenuators, amplifiers and undercutters of N and N' are identical (that is $A_{\delta}^{-}(N) = A_{\delta}^{-}(N')$, $A_{\delta}^{+}(N) = A_{\delta}^{+}(N')$, $A_{\delta}^u(N) = A_{\delta}^u(N')$) and have the same final weights in both graphs (that is, for every $\delta^* \in A_{\delta}^u(N) \cup A_{\delta}^{+}(N) \cup A_{\delta}^{-}(N)$, $f_N^w(\delta^*) = f_{N'}^w(\delta^*)$), then $f_N^w(\delta) = f_{N'}^w(\delta)$.*

The next principle, Independence of Modifiers, is a stronger version of Immediate Successors. It says, in effect, that the types of modifiers do not interact in what they contribute to amplifying, attenuating or undercutting a consideration. We could, for example, add up their (final) weight individually first and then use their joint weight to determine the final weight of the consideration they modify (or undercut).

Principle 3.4.7 (Independence of Modifiers). *There is a function s from a (multi-)set of real numbers to the non-negative real numbers, such that given a Normative Graph N and a weighting function w where for some consideration δ of N , let M^{+} be the (multi-)set of final weights of A_{δ}^{+} , M^{-} the (multi-)set of final weights of A_{δ}^{-} and M^u the (multi-)set of final weights of A_{δ}^u . Then, $f_N^w(\delta)$ can be determined solely based on $w(\delta)$, $s(M^{+})$, $s(M^{-})$, and $s(M^u)$.*

When Independence of Modifiers holds, we can define the next principle, Weak Modifier Reciprocity, which says that if the joint final weight of the amplifiers is equal to the joint weight of the attenuators of some considerations, then the final weight of that consideration is equal to its default weight as long as it is not undercut. This is thus a stronger version of the principle Unmodified Default.

Principle 3.4.8 (Weak Modifier Reciprocity). *This principle is only well-defined if Principle Independence of Modifiers holds. Let s be a function as described in Principle Independence of Modifiers. Let N be an arbitrary Normative Graph. Then if for some such function s , if for a given consideration δ , $s(M^{+}) = s(M^{-})$ and there is no undercutter of δ with a positive final weight, then $f_N^w(\delta) = w(\delta)$.*

We introduce another version of a principle capturing the reciprocity of modifiers (independently of whether Principle Independence of Modifiers holds.) The next principle, Strong Modifier Reciprocity, states that if a consideration is modified by an amplifier and an attenuator of equal weight, then removing both of them does not affect the final weight of the consideration. That is, amplifiers and attenuators *cancel out*.²⁴

²⁴Even though we have called the two principles “weak” and “strong”, the weak version does not follow from the strong version even if well-defined. Still, it strikes us that the “strong” version is the more general one.

3.4. A Principle-Based Analysis of Reasons and Their Final Weight

Principle 3.4.9 (Strong Modifier Reciprocity). *If, given a weighting function w and Normative Graph N , δ^- is an attenuator of δ and δ^+ is an amplifier of δ with $f_N^w(\delta^-) = f_N^w(\delta^+)$, then $f_N^w(\delta) = f_{N-\{\delta^+, \delta^-\}}^w(\delta)$.*

The next principle, No Enablers, posits that so-called enablers do not exist. If a consideration has a default weight of 0, then (no matter the context) the final weight of the consideration is 0 as well.²⁵

Principle 3.4.10 (No Enablers). *For any weighting function w and Normative Graph $N = (\Delta, O)$, if for some consideration $\delta \in \Delta$, $w(\delta) = 0$, then $f_N^w(\delta) = 0$.*

The next principle, Relativity – Final Weights, says that the final weights of a consideration capture only a relative difference.²⁶

Principle 3.4.11 (Relativity – Final Weights). *For any positive real number k , weighting function w , Normative Graph N , for all considerations δ of N , $k * f_N^w(\delta) = f_N^{w*k}(\delta)$.*

Principle Distinct Roles guarantees that no attenuator (or amplifier, for that matter) ever undercuts a consideration completely.

Principle 3.4.12 (Distinct Roles). *Given a consideration δ in an arbitrary Normative Graph N with $w(\delta) > 0$. If for every undercutter δ^u of δ , $f_N^w(\delta^u) = 0$, then $f_N^w(\delta) > 0$.*

This principle, the converse of Attenuators as Undercutters, says that any consideration can be undercut by a strong enough attenuator (or a set thereof). Note that the two principles Distinct Roles and Attenuators as Undercutters are contraries, but not contradictories. It is possible for neither to hold.

Principle 3.4.13 (Attenuators as Undercutters). *Given a Normative Graph $N = (\Delta, O)$, a weighting function w , and a consideration $\delta \in \Delta$, there is a finite set of considerations X , disjoint from Δ , and a weighting function w^* such that for all $\delta' \notin X$, $w(\delta') = w^*(\delta')$, where for all $\delta^- \in X$, $\text{Conclusion}(\delta^-) = \delta$ and $\text{Role}(\delta^-) = a^-$, such that $f_{N+X}^{w^*}(\delta) = 0$.*

Principle F-Restriction states that if all default weights are rational (alternatively natural) numbers, then all final weights are rational (alternatively natural) numbers.

²⁵We take the word *enabler* from (Dancy, 2004). Dancy might not agree with our description of what an enabler is. In any case, for present purposes, we treat it as something that, if it is present, makes a reason have positive weight and if it is absent makes it have a weight of 0.

²⁶This principle will turn out very plausible for reasons, but not so for determining their final weights.

3. The Kitchen Scale Model Formalized

Principle 3.4.14 (F-Restriction). *If for every consideration δ of some Normative Graph N , $w(\delta) \in \mathbb{Q}$ (analogously $\mathbb{N}$), then for all considerations δ , $f_N^w(\delta) \in \mathbb{Q}$ (analogously $\mathbb{N}$).*

The next principle, Atemporality, needs some unpacking. It says that the calculation of the final weight can be done iteratively. One can calculate the final weight of a consideration by first calculating the final weight it would have given only a subset of its amplifiers/attenuators etc., then treat that as its new default weight and calculate the final weight given the remaining modifiers and this new default weight. One can think of this in two ways. First, this models the temporal dimension of new considerations popping up. Atemporality says that one can simply update the weights based solely on the new considerations alone, without regard of what other considerations were already in place. Second, one can view it as a statement about the contribution each consideration makes in the context of others. Atemporality guarantees that each contribution to the final weight of a consideration is independent. If Atemporality is violated, then the contribution of a consideration depends on what other considerations are in play. Atemporal ways to calculate final weights are therefore more on the side of atomism.

Principle 3.4.15 (Atemporality). *Let $N = (\Delta, O)$ be a Normative Graph and w a weighting function. Given a consideration $\delta \in \Delta$, let A_δ^+ be the set of its amplifiers, A_δ^- be the set of its attenuators and A_δ^u the set of its undercutters. Let $\Delta^* \subset A_\delta^+ \cup A_\delta^- \cup A_\delta^u = \Delta'$. Let w^* be identical to w except that $w^*(\delta) = f_{N-\Delta^*}^w(\delta)$, then $f_N^w(\delta) = f_{N-(\Delta' \setminus \Delta^*)}^{w^*}(\delta)$.*

The next set of principles cover whether quality or quantity is more important when calculating the final weights of considerations.

The first of these principles, Modifier Scaling: Equality, states that in any Normative Graph, n modifiers of the same type and the same default weight can be replaced by a single modifier of that type with exactly n times their default weight without changing any final weights.

Principle 3.4.16 (Modifier Scaling: Equality). *Given a Normative Graph $N = (\Delta, O)$, a weighting function w , and n modifiers $\delta_1, \dots, \delta_n \in \Delta$, with $\text{Role}(\delta_1) = \dots = \text{Role}(\delta_n)$, $w(\delta_1) = \dots = w(\delta_n)$ and $\text{Conclusion}(\delta_1) = \dots = \text{Conclusion}(\delta_n) = \delta$, let $\delta^* \notin \Delta$ with $\text{Role}(\delta^*) = \text{Role}(\delta_1)$, $\text{Conclusion}(\delta^*) = \delta$ and $w(\delta^*) = n * w(\delta_1)$, then $f_N^w(\delta) = f_{N-\{\delta_1, \dots, \delta_n\}+\delta^*}^w(\delta)$.*

In contrast, Modifier Scaling: Quality says something to the effect that the quality of modifiers is more important than the quantity. In any Normative Graph, replacing n modifiers of the same type and the same default weight with a single modifier of that type and n times their default weight, leads to the final weight of the modified consideration being at least as large as it was before.

3.4. A Principle-Based Analysis of Reasons and Their Final Weight

Principle 3.4.17 (Modifier Scaling: Quality). *Given a Normative Graph $N = (\Delta, O)$, a weighting function w and n modifiers $\delta_1 \dots, \delta_n \in \Delta$, with $\text{Role}(\delta_1) = \dots = \text{Role}(\delta_n)$, $w(\delta_1) = \dots = w(\delta_n)$ and $\text{Conclusion}(\delta_1) = \dots = \text{Conclusion}(\delta_n) = \delta$, let $\delta^* \notin \Delta$ with $\text{Role}(\delta^*) = \text{Role}(\delta_1)$, $\text{Conclusion}(\delta^*) = \delta$ and $w(\delta^*) = n * w(\delta_1)$, then $f_N^w(\delta) \leq f_{N-\{\delta_1, \dots, \delta_n\}+\delta^*}^w(\delta)$.*

Analogously, Modifier Scaling: Quantity says that quantity trumps quality. In any Normative Graph, replacing n modifiers of the same type and the same default weight with a single modifier of that type and n times their default weight, leads to the final weight of the modified consideration being at most as large as it was originally.

Principle 3.4.18 (Modifier Scaling: Quantity). *Given a Normative Graph $N = (\Delta, O)$, a weighting function w and n modifiers $\delta_1 \dots, \delta_n \in \Delta$, with $\text{Role}(\delta_1) = \dots = \text{Role}(\delta_n)$, $w(\delta_1) = \dots = w(\delta_n)$ and $\text{Conclusion}(\delta_1) = \dots = \text{Conclusion}(\delta_n) = \delta$, let $\delta^* \notin \Delta$ with $\text{Role}(\delta^*) = \text{Role}(\delta_1)$, $\text{Conclusion}(\delta^*) = \delta$ and $w(\delta^*) = n * w(\delta_1)$, then $f_N^w(\delta) \geq f_{N-\{\delta_1, \dots, \delta_n\}+\delta^*}^w(\delta)$.*

With these principles in place, we move on to the next stage of the principle-based analysis. Categorizing functions according to whether they violate or satisfy each principle. In the next section, we will introduce four functions that we will classify in such a way.

Functions

In this section, then, we will introduce four functions $f_1, \dots, f_4$, as candidates to calculate the final weights. As, so far, there is little discussion in the philosophical literature how precisely the final weight of a reason (or some other consideration) is determined, we will not make any claims to have found the final answer here. Instead, we aim to show what such functions could look like by viewing them through the principles introduced above. We consider it entirely possible that another, better, function will be found, and we also think that the principles can serve as a guide for identifying potential better functions.

We have already defined a shorthand for undercutters, that we repeat here.

Definition 3.4.5 (Undercutter Shorthand). *Let δ be some consideration in some Normative Graph N , A_δ^u be the set of undercutters of δ , let $f \in \{f_1, \dots, f_4\}$, then we define²⁷:*

$$u(\delta) = \begin{cases} 1 & \text{if } \sum_{\delta' \in A_\delta^u} f(\delta') = 0 \\ 0 & \text{otherwise} \end{cases}$$

²⁷As f depends on w and N , so does u . However, to make the notation less confusing, we omit explicitly mentioning w and N in the definition of u .

3. The Kitchen Scale Model Formalized

Note, again, that as our graphs are acyclic, this function, and all the ones we introduce below, are well-defined. We will be able to calculate the final weight of each consideration in a Normative Graph from the leaves up. To keep the functions tractable, we suppress the reference to the Normative Graph, wherever no confusion can arise.

The first function we have already seen. It is the *Additive Function*. It states that what amplifiers do is add their weight to that of a reason, while attenuators subtract their weight from it. Since we do not want the weight of a normative consideration to become negative (see Principles No Valence Flips for Modifiers and No Valence Flips for Reasons), we set the minimal weight to 0. As with all the other functions, we also set the weight to 0 if the consideration is undercut.

Definition 3.4.6 (Additive Function).

$$f_1(\delta) = \max(0, w(\delta) + \sum_{a^+ \in A_\delta^+} f_1(a^+) - \sum_{a^- \in A_\delta^-} f_1(a^-)) * u(\delta)$$

What distinguishes the Additive Function from the following ones is that it is unique among the function we propose in violating principle Distinct Roles and No Enablers. The simple Additive Function allows attenuators, if they are strong enough, to completely take away the weight of a consideration. Conversely, it allows for *inactive* considerations, those with a default weight of 0, to have a positive final weight, by way of amplifiers.

As it is purely additive, it also has a feature that might seem counterintuitive. A very weak consideration, a reason, say, can become very strong if it is amplified by a very strong amplifier. That means, the default weight is not as important. One might assume, but we do not do so here, that a strong amplifier and a weak default weight combine to a medium final weight. That is what the next function does.

Instead of summing the final weights of amplifiers and subtracting attenuators, the next function multiplies the default weight by 1 plus the weight of each amplifier and divides by 1 plus the weight of each attenuator. Adding the 1 guarantees that a consideration with a default weight of 0 does not affect the final weight (see Principle No Effects from Spurious Modifiers).²⁸

²⁸This is somewhat close to the only explicitly suggested functions mentioned in the literature we are aware of due to (Bader, 2016). Bader mentions in passing that modifiers work akin to multiplying a reason's strength with some number, (amplifiers with a number larger than 1, attenuators with a number less than 1). This simplistic approach has, we think, a number of unintuitive consequences. For one, a modifier of strength 0 is not weak, but the strongest attenuator it could be. Second, and more importantly, it violates what we take to be intuitive assumptions about nested modifiers. For example, a (strongly) amplified attenuator can become itself an amplifier and not, as we think it should be, stay an even stronger attenuator.

3.4. A Principle-Based Analysis of Reasons and Their Final Weight

Definition 3.4.7 (Multiplicative Function).

$$f_2(\delta) = w(\delta) * \frac{\prod_{a^+ \in A_\delta^+} (1 + f_2(a^+))}{\prod_{a^- \in A_\delta^-} (1 + f_2(a^-))} * u(\delta)$$

To see this function in action, let us apply it to our toy example as well. As a reminder, our Normative Graph comprised the following considerations (to which we added our weights):

- $\delta_1 = (\textit{Promise}, \textit{EatCake}, r^+)$ with $w(\delta_1) = 3$
- $\delta_2 = (\textit{Hungry}, \textit{GiveCake}, r^+)$ with $w(\delta_2) = 4$
- $\delta_3 = (\textit{TakePleasureYou}, \textit{EatCake}, r^+)$ with $w(\delta_3) = 2$
- $\delta_4 = (\textit{TakePleasureNiece}, \textit{GiveCake}, r^+)$ with $w(\delta_4) = 2$
- $\delta_5 = (\textit{StoleAToy}, \textit{GiveCake}, r^-)$ with $w(\delta_5) = 3$
- $\delta_6 = (\textit{Regrets}, \delta_5, a^-)$ with $w(\delta_6) = 2$
- $\delta_7 = (\textit{Release}, \delta_1, u)$ with $w(\delta_7) = 4$

Like for the Additive Function, we start by calculating final weights of the leaves of the graph. Their final weight is just their default weight. This means that $f_2(\delta_2) = 4, f_2(\delta_3) = 2, f_2(\delta_4) = 2, f_2(\delta_6) = 2, f_2(\delta_7) = 4$. All that is left is to calculate the final weight of δ_1 and δ_5 . But δ_1 is undercut by δ_7 and so $f_2(\delta_1) = 0$. For the final weight of δ_5 , we need to divide $w(\delta_5) = 3$ by $1 + f_2(\delta_6) = 2$; so here too we get $f_2(\delta_5) = 1$.

This function violates a principle that all others satisfy: Modifier Scaling: Equality. It states that n modifiers of the same weight (and type) contribute the same as a single modifier of n times their weight. Clearly, if one multiplies the numbers, this is no longer the case. We believe that a function that does satisfy Modifier Scaling: Equality is preferable.

The next function tries to remedy this. Instead of multiplying/dividing by the weight of each modifier, we sum the weights of amplifiers and multiply by 1 plus their sum and then divide by 1 plus the sum of the attenuators.

Definition 3.4.8 (Mixed Function I).

$$f_3(\delta) = w(\delta) * \frac{1 + \sum_{a^+ \in A_\delta^+} f_3(a^+)}{1 + \sum_{a^- \in A_\delta^-} f_3(a^-)} * u(\delta)$$

3. The Kitchen Scale Model Formalized

We can easily see that in our toy example, the final weights of f_3 are the same as those of f_2 , as δ_1 remains undercut, and we again have to divide $w(\delta_5)$ by $1 + f_3(\delta_6) = 1 + 2$. So again, $f_3(\delta_5) = 1$.

This function satisfies (what we take) to be a good set of principles. Notably, though, it violates Strong Modifier Reciprocity, that is, attenuators and amplifiers of equal strength do not cancel out. We might want this principle and keep the other principles as they are. The next function will accomplish that.

Instead of multiplying by (1 plus) the sum of amplifiers and dividing by (1 plus) the sum of the attenuators, we multiply by (1 plus) their difference if their difference is positive. If it is negative, we divide by (the absolute value of) (1 plus) their difference.

Definition 3.4.9 (Mixed Function II).

$$f_4(\delta) = \begin{cases} w(\delta) * (1 + \sum_{a^+ \in A_\delta^+} f_4(a^+) - \sum_{a^- \in A_\delta^-} f_4(a^-)) * u(\delta) & \text{if } \sum_{a^+ \in A_\delta^+} f_4(a^+) \geq \sum_{a^- \in A_\delta^-} f_4(a^-) \\ w(\delta) * \frac{1}{(1 + \sum_{a^- \in A_\delta^-} f_4(a^-) - \sum_{a^+ \in A_\delta^+} f_4(a^+))} * u(\delta) & \text{otherwise} \end{cases}$$

f_4 gives, as can be seen easily, the same results for all final weights in our toy example as f_3 . And so, in our toy example, all four functions agree on the final weight. This is, of course, simply due to the construction of the example and the choice of the weights we put on the considerations and in general, the functions will give very different results for the final weights of normative considerations in different scenarios.

We will not exclusively commit to any of the functions we have proposed here. Our intuitions align most closely with the set of principles satisfied for the last function.²⁹ Still, we want to refrain from committing to either any particular function or any particular set of principles.

Before we move on, let us quickly take stock of what is – in principle – possible with the principles we introduced. We can show that holding fast to certain combinations of them is untenable. For example, we can show that some combination of principles does not allow for any non-trivial (rational) function to calculate the final weights.

Proposition 3.4.1. *Every rational weighting function that satisfies the following principles is a constant function assigning 0 to every consideration in any Normative Graph.*

²⁹Save, perhaps, for the fact that it violates Atemporality, which we are not sure about.

3.4. A Principle-Based Analysis of Reasons and Their Final Weight

1. *Modifier Scaling: Equality*
2. *No Valence Flips for Modifiers*
3. *Strong Modifier Reciprocity*
4. *No Enablers*
5. *Attenuators as Undercutters*
6. *Atemporality*

This is a kind of impossibility result, but, to be clear, it does not nearly reach the generality of Arrow’s impossibility theorem. We include this proposition as a proof of concept. We think it is fruitful to reason about principles in this way and exclude certain combinations of principles, but seeing what (further) principles lead to inconsistencies is left for future work. The proof of the above proposition can be found in Appendix A.3.

Principle	Additive	Multiplicative	Mixed I	Mixed II
Anonymity	yes	yes	yes	yes
Unmodified Default	yes	yes	yes	yes
Modeler’s Delight	yes	yes	yes	yes
No Effects from Spurious Modifiers	yes	yes	yes	yes
No Valence Flips for Modifiers	yes	yes	yes	yes
Immediate Successors	yes	yes	yes	yes
Independence of Modifiers	yes	yes	yes	yes
Weak Modifier Reciprocity	yes	yes	yes	yes
Strong Modifier Reciprocity	yes	yes	no	yes
No Enablers	no	yes	yes	yes
Relativity – Final Weights	yes	no	no	no
Distinct Roles	no	yes	yes	yes
Attenuators as Undercutters	yes	no	no	no
F-Restriction	$\mathbb{N}, \mathbb{Q}$	$\mathbb{Q}$	$\mathbb{Q}$	$\mathbb{Q}$
Atemporality	no	yes	no	no
Modifier Scaling: Equality	yes	no	yes	yes
Modifier Scaling: Quality	yes	For attenuators	yes	yes
Modifier Scaling: Quantity	yes	For amplifiers	yes	yes

Table 3.1.: Overview of which principles the functions $f_1, \dots, f_4$ satisfy. Proofs are simple, but still included in Appendix A.3.

Let’s sum up what we have achieved so far: In Section 3.3.1, we have put in place the structure of a formal model of weighing reasons through the notion of Normative Graphs, and shown with some simple functions how we can derive final weights and

3. The Kitchen Scale Model Formalized

deontic statuses. Next, in Section 3.4.1, we have analyzed in depth what constraints we might want to put on functions that determine the final weights of reasons (and other considerations) and introduced four candidates for such functions. What is left to do is to describe how we can go from the final weights of reasons to the actions we are permitted, forbidden, or required to do in a principled manner.

3.4.2. Weighing Reasons

So in this section, we will answer the question of how we determine the deontic statuses of actions, once the final weights of reasons are in place. We can consider this to be first aggregating the joint weight of reasons and then weighing it. Here too we will undertake a principle-based analysis, but there are two differences; first we're going to keep our analysis on the second stage and not prove any kind of determination or impossibility result whatsoever. Second, unlike in the case for functions determining the final weight, there actually are some explicit and well-developed accounts found in the philosophical literature on how to determine the deontic statuses based on the final weights of reasons, the most comprehensive of which is probably (Tucker, forthcominga).

What we are after is a function g , that, given a Normative Graph $N = (\Delta, O)$, a weighting function w and a function $f \in \{f_1, \dots, f_4\}$ returns a subset of O – the set of permissible actions in N – and does so based on the final weights of reasons in N (according to f). We will call this function g , either an *aggregating* function or a *weighing* function, as – with the kitchen scale model as background – what the function is doing is akin to aggregating the joint weight of reasons for and against an action and then weighing it against competing options.³⁰

The choice of g is largely conceptually independent of the choice of f and any combination of the two will give a unique semantics.³¹ Something we will explore in Section 3.5.1. First, however, we need to introduce some further notation.

Definition 3.4.10 (Reasons for an option). *Let $N = (\Delta, O)$ be a Normative Graph and let $o \in O$. We then define:*

$$\bullet R_o^+(N) = \{\delta \in \Delta : \text{Conclusion}(\delta) = o \text{ and Role}(\delta) = r^+\}$$

³⁰We will introduce some functions that arguably do not conform well to this metaphor, but we believe the labels for these functions are still descriptive.

³¹Of course, this model could be generalized easily. Instead of having a modular approach that calculates the final weight of reasons first and then – based solely on those reasons and their final weights – determines the normative status of actions, one could more broadly define a function from a Normative Graph $N = (\Delta, O)$ to a subset of O without having it be the composition of two modular functions. We don't believe there is either precedent found in the philosophical literature for such an approach, nor is it independently plausible.

3.4. A Principle-Based Analysis of Reasons and Their Final Weight

- $R_o^-(N) = \{\delta \in \Delta : \text{Conclusion}(\delta) = o \text{ and } \text{Role}(\delta) = r^-\}$
- $R_N = \bigcup_{o \in O} (R_o^+ \cup R_o^-)$.

Wherever the context makes it clear, which Normative Graph we are talking about, we omit mentioning it explicitly.

Then, for a Normative Graph $N = (\Delta, O)$, what we are after is a function $g_N^f : R_N \rightarrow \mathcal{P}(O)$. Two things are of note here. First, as f depends on a weighting function w , g_N^f , too is implicitly defined in terms of w . However, the concrete choice of w , and even f , is usually not relevant to stating any of the principles we discuss below, and so we will frequently omit it. The same goes for the Normative Graph in the subscript. In fact, the notation of R_N mentions the graph and in turn the reasons R_N can be derived from the Normative Graph N . So we will often write $g(R_N)$ to mean $g_N^f(R_N)$ for some function f .

So let's get started with the principle-based analysis.

Principles

Like before, we will give a natural-language gloss of the principle before giving its formal definition. After we have introduced the set of principles, we will again propose concrete functions for g . We give a compact overview of which principles are satisfied or violated by the various functions in Table 3.2.

Our first principle states that, all things equal, every action is permissible. If there are no reasons for or against any action, they are all permissible.³²

Principle 3.4.19 (All Allowed). *Given a Normative Graph $N = (\Delta, O)$ with $R_N = \emptyset$, $g(R_N) = O$.*

The next principle, Anonymity I, makes sure that the *names* of reasons do not make a difference in the deontic status of actions but only their final weights.

Principle 3.4.20 (Anonymity I). *Let $N = (\Delta, O)$ and $N' = (\Delta', O)$. We will call R_N and $R_{N'}$ structurally identical, if there is a bijection ϕ from R_N to $R_{N'}$ such that*

³²This contrasts with the epistemic case where the standard position is that if one has equally strong reasons for believing some statement and its negation, one is not permitted to either believe the statement or believe that the statement is false, but rather withhold belief the statement entirely.

3. The Kitchen Scale Model Formalized

for all $r \in R$, $\text{Conclusion}(r) = \text{Conclusion}(\phi(r))$, $\text{Role}(r) = \text{Role}(\phi(r))$ and $f_N^w(r) = f_{N'}^w(\phi(r))$,

then for any structurally identical sets of reasons R_N and $R_{N'}$, $g(R_N) = g(R_{N'})$.

The second of the two anonymity principles, Anonymity II, states that renaming the options does not change the outcome.

Principle 3.4.21 (Anonymity II). *Let $N = (\Delta, O)$, $N' = (\Delta', O')$. Let there be a bijection ϕ between O and O' and a bijection ψ between R_N and $R_{N'}$. Let R_N be such that for every $o \in O$ and $r \in R_N$, $\text{Conclusion}(r) = o$ just in case $\text{Conclusion}(\psi(r)) = \phi(o)$ and additionally, $\text{Premise}(r) = \text{Premise}(\psi(r))$ and $f(r) = f(\psi(r))$. Then $O \subseteq g(R_N)$ iff $\phi(O) \subseteq g(R_{N'})$.*

The next principle, No Effects from Spurious Reasons, says that adding reasons with (final) weight 0 has no impact on which actions are permissible and which are not.

Principle 3.4.22 (No Effects from Spurious Reasons). *Let w be an arbitrary weighting function, $N = (\Delta, O)$ be a Normative Graph and $r \notin \Delta$. Let $f_{N+r}^w(r) = 0$. Then $g(R_N) = g(R_{N+r})$.*

The next principle, No Valence Flips for Reasons, states that adding another reason in favour does not make an action suddenly impermissible.³³

Principle 3.4.23 (No Valence Flips for Reasons). *Let $N = (\Delta, O)$ be a Normative Graph, and r a reason. If $\text{Conclusion}(r) = o \in O$ and $\text{Role}(r) = a^+$ (but r is not necessarily in Δ) and $o \in g(R_N)$, then $o \in g(R_{N+r})$. Analogously, if $\text{Conclusion}(r^*) = o^*$ and $\text{Role}(r^*) = r^-$ and $o^* \notin g(R_N)$, then $o^* \notin g(R_{N+r^*})$.*

Similar to Independence of Modifiers, the next principle, Independence of Reasons, says that normative statuses can be derived by first aggregating the individual reasons for or against.

Principle 3.4.24 (Independence of Reasons). *Let M_o^+ name the (multi-)set of final weights of R_o^+ for some action o in some Normative Graph N ; Analogously, let M_o^- be the (multi-)set of final weights of R_o^- . There is a function s from these sets to the real numbers, such that for every Normative Graph $N = (\Delta, O)$; $g(R_N)$ can be derived solely from the combination of $s(M_o^+)$ and $s(M_o^-)$ (for each option o).*

³³Note that this only holds on the level of reasons. Adding some fact that is both a reason for an action and an attenuator or undercutter for some other reason for that action might well change the action's normative status. We will discuss this principle and cases like this in more detail in Section 3.5.1.

3.4. A Principle-Based Analysis of Reasons and Their Final Weight

If Independence of Reasons holds, we can define the next principle, Weak Reason Reciprocity, which states that aggregates of reasons for and reasons against cancel out.

Principle 3.4.25 (Weak Reason Reciprocity). *If the principle Independence of Reasons holds, and for some such function s and sets defined above, $s(M_o^+) = s(M_o^-)$, then $g(R_N) = g(R_{N-(R_o^+(N) \cup R_o^-(N))})$.*

Like in the case of modifiers, we can define a strong version of this principle as well: Strong Reason Reciprocity. It says that even individual reasons for and against cancel out.

Principle 3.4.26 (Strong Reason Reciprocity). *If $r \in R_o^+(N)$ and $r^* \in R_o^-(N)$ with $f_N^w(r) = f_N^w(r^*)$, then $g(R_N) = g(R_{N-\{r, r^*\}})$.*

The next principle, Relativity – Reasons, has some precedent in the literature, even though it is usually not argued for (see (Drai, 2018)). It says that the final weight of reasons is a purely relative measure.

Principle 3.4.27 (Relativity – Reasons). *Let $N = (\Delta, O)$ and $N' = (\Delta', O)$, such that there is a bijection ϕ from R_N to $R_{N'}$ and for every $r \in R_N$: $\text{Conclusion}(r) = \text{Conclusion}(\phi(r))$, $\text{Role}(r) = \text{Role}(\phi(r))$ and $f_N^w(r) = k * f_{N'}^w(\phi(r))$, then $g(R_N) = g(R_{N'})$.*

The next principle, No Dilemmas states that there are no genuine normative dilemmas; that is, in every situation, *some* action is permitted.

Principle 3.4.28 (No Dilemmas). *Let N be an arbitrary Normative Graph, then $g(R_N) \neq \emptyset$*

The next two principles, like the principles Modifier Scaling: Quality, Modifier Scaling: Equality, Modifier Scaling: Quantity in the case of modifiers, measure whether quality or quantity is more important when it comes to reasons.

Reasons Scaling: Quality says that a single reason that is n times as strong as each of n reasons, will work toward making an action permissible at least as well the n reasons.

Principle 3.4.29 (Reasons Scaling: Quality). *Let $N = (\Delta, O)$ be an arbitrary Normative Graph containing n reasons $r_1, \dots, r_n \in R_N$ with $\text{Conclusion}(r_1) = \dots =$*

3. The Kitchen Scale Model Formalized

Conclusion(r_n) = o and $f_N^w(r_1) = \dots = f_N^w(r_n)$. Let r^* be a reason with *Conclusion*(r^*) = o not in Δ , such that $f_{N+r^*}^w(r^*) = n * f_N^w(r_1) = \dots = n * f_N^w(r_n)$. If $o \in g(R_N)$, then $o \in g(R_{N-\{r_1, \dots, r_n\}+r^*})$.

Reasons Scaling: Quantity is the inverse of Reasons Scaling: Quality and says that n reasons of some uniform strength will work at least as well as a reason of n times that strength in making an action permissible.

Principle 3.4.30 (Reasons Scaling: Quantity).

Let $N = (\Delta, O)$ be an arbitrary Normative Graph containing n reasons $r_1, \dots, r_n \in R_N$ with *Conclusion*(r_1) = $\dots$ = *Conclusion*(r_n) = o and $f_N^w(r_1) = \dots = f_N^w(r_n)$. Let r^* be a reason with *Conclusion*(r^*) = o not in Δ , such that $f_{N+r^*}^w(r^*) = n * f_N^w(r_1) = \dots = n * f_N^w(r_n)$. If $o \in g(R_{N-\{r_1, \dots, r_n\}+r^*})$, then $o \in g(R_N)$.

The next principle, Difference Maker, says that, for every (potential) reason, no matter how weak, there is a scenario where it makes a difference as to what to do.

Principle 3.4.31 (Difference Maker). Given a reason r in favour of o and a reason r^* against o , there is a weighting function w , and Normative Graphs N, N' with $f_{N+r}^w(r) \neq 0$ and $f_{N'+r^*}^w(r^*) \neq 0$, such that $o \notin g(R_N)$ and $o \in g(R_{N+r})$ and $o \in g(R_{N'})$ and $o \notin g(R_{N'+r^*})$.

For the next two principles, we will introduce some new notation. The next definition makes precise, what it means to change the polarity of a reason.

Definition 3.4.11 (Polarity Swapping). Given a reason $r = (p, o, x)$, r_+ is the reason (p, o, r^+) . Given a set of reasons R , the set R_+ is the set $\{r_+ | r \in R\}$.

We define r_- and the set R_- analogously with reasons against instead of reasons in favour of.

The next definition makes precise, what it means to change the action a reason is a reason for or against.

Definition 3.4.12 (Option Swapping). Given a reason $r = (p, o^*, x)$ and some option $o \in L$, r^o is the reason (p, o, x) . Given a set of reasons R , the set R^o is the set $\{r^o | r \in R\}$.

With these definitions, we can define the next principle – Reasons Symmetry I. It states, in effect, that reasons against some option can simply be treated as reasons for all other

3.4. A Principle-Based Analysis of Reasons and Their Final Weight

options. This has the immediate effect that reasons against can in principle be eliminated entirely.³⁴

Principle 3.4.32 (Reasons Symmetry I). *For some Normative Graph N , let $R_N = R_{o_1}^+ \cup R_{o_1}^- \cup \dots \cup R_{o_n}^+ \cup R_{o_n}^-$. Let, for some Normative Graph N' , $R_{N'} = R_{o_1}^+ \cup (R_{o_2}^+ \cup R_{o_2}^- \cup (R_{o_1}^+)^{o_2}) \cup \dots \cup (R_{o_n}^+ \cup R_{o_n}^- \cup (R_{o_1}^+)^{o_n})$ such that the final weights of the reasons (and their polarity/option swapped versions) in N and N' coincide, then $g(R_N) = g(R_{N'})$.*

Principle Reasons Symmetry II is the converse of Reasons Symmetry I. It states that reasons in favour can be treated as reasons against other options.

Principle 3.4.33 (Reasons Symmetry II). *For some Normative Graph N , let $R_N = R_{o_1}^+ \cup R_{o_1}^- \cup \dots \cup R_{o_n}^+ \cup R_{o_n}^-$. Let, for some Normative Graph N' , $R_{N'} = R_{o_1}^- \cup (R_{o_2}^+ \cup R_{o_2}^- \cup (R_{o_1}^+)^{o_2}) \cup \dots \cup (R_{o_n}^+ \cup R_{o_n}^- \cup (R_{o_1}^+)^{o_n})$ such that the final weights of the reasons (and their polarity/option swapped versions) in N and N' coincide, then $g(R_N) = g(R_{N'})$.*

The last principle, Menu Insensitivity, states that adding more options doesn't make a permissible option impermissible.

Principle 3.4.34 (Menu Insensitivity). *Given two Normative Graphs $N = (\Delta, O)$ and $N' = (\Delta, O')$ with $O \subseteq O'$, $g(R_N) \subseteq g(R_{N'})$.*

With this, we conclude introducing principles for weighing reasons based on their final weights. Like we mentioned at the beginning of this section, for the principles governing the weights of reasons, we confine our analysis to stage two of the principle-based analysis and will not provide either a full characterization of some functions in terms of principles or prove that some (interesting) combination of principles cannot be satisfied by any function. Instead, like we did for the functions governing the final weights of considerations, we will introduce several contender functions and then compare them through the lens of the principles.

Functions

We are looking for a function g that, for each Normative Graph given some function f determining the final weights of reasons, returns a subset of the possible actions that we

³⁴Of course, this only concerns the formal level. In terms of *truly representing the normative world* it might well be that reasons against are distinct. See (Snedegar, 2018) for a discussion of whether reasons against can be eliminated. It should be noted that the conclusion he reaches is one decidedly rejected here.

3. The Kitchen Scale Model Formalized

are permitted to perform. Additionally, we say that an option is required, if it is the only permissible option and forbidden, if it is not permissible. As before, we suppress the weighting function w and the Normative Graph in our notation. For now, we assume the function governing the final weight of the reasons f is one of the functions $f_1, \dots, f_4$ we introduced above. Throughout, we assume our normative graph to be $N = (\Delta, O)$.

The first function we have already seen. We originally introduced it in Section 3.3.2 due to its simplicity. However, it is also the function that is most often referred to in the non-formal literature (in fact, it is the only function an analogue of which is defended in an entire book, (Tucker, forthcominga)). It simply states that an action is permissible if, and only if, the sum of the (weights of the) reasons for it minus the sum of the (weights of the) reasons against it is larger than this difference for any other option.

Definition 3.4.13 (Sum of Reasons).

$$g_1(R_N) = \{o \in O \mid \text{for all } o^* \in O : \sum_{r \in R_o^+} f(r) - \sum_{r \in R_o^-} f(r) \geq \sum_{r \in R_{o^*}^+} f(r) - \sum_{r \in R_{o^*}^-} f(r)\}$$

Function g_1 is unique among the functions we propose in that it satisfies almost every principle we have introduced. The only principle not satisfied is Menu Insensitivity. We think this partly explains the popularity of this simple function in the literature.

The second function makes formal the idea that we are only required to do *well enough*. That is, there is a threshold (represented by some real number) that the (weight of) our reasons for and against have to satisfy, but as long as they do, everything else is optional.

Definition 3.4.14 (Additive Threshold). *Let k be some fixed real number.*

$$g_2(R_N) = \{o \in O \mid \sum_{r \in R_o^+} f(r) - \sum_{r \in R_o^-} f(r) \geq k\}$$

There is a special case of the function, $k = 0$, that is worth mentioning explicitly. Here, you simply have to consider the reasons for and against any single action to see if it is permissible.

Definition 3.4.15 (Solipsistic Reasons).

$$g_3(R_N) = \{o \in O \mid \sum_{r \in R_o^+} f(r) \geq \sum_{r \in R_o^-} f(r)\}^{35}$$

³⁵We choose this name here, because whether one is permitted to do an action does not depend on (the existence of) any other actions one could do instead.

3.4. A Principle-Based Analysis of Reasons and Their Final Weight

Both g_2 and g_3 satisfy Menu Insensitivity and violate No Dilemmas. This is because the way the deontic status of an action is derived is completely independent of other actions and the reasons for them. This means (unsurprisingly) that the symmetry principles Reasons Symmetry I and Reasons Symmetry II are violated. Still, all other principles are satisfied.

Next, we introduce a function that excludes reasons if they are too weak. It aims to model the idea that minuscule factors, like one's preference for chocolate ice cream over vanilla, are not the kind of factors that generate genuine normative requirements.

Definition 3.4.16 (Sum of Threshold-Reasons). *Let $l \geq 0$ be some fixed real number.*

$$\begin{aligned}
 g_4(R_N) = \{o \in O \mid & \text{for all } o^* \in O : \\
 & \sum_{r \in R_o^+ : f(r) > l} f(r) - \sum_{r \in R_o^- : f(r) > l} f(r) \\
 & \geq \\
 & \sum_{r \in R_{o^*}^+ : f(r) > l} f(r) - \sum_{r \in R_{o^*}^- : f(r) > l} f(r)\}
 \end{aligned}$$

This function is obviously identical to g_1 if $l = 0$ for every $f \in \{f_1, \dots, f_4\}$. It is also (together with the very similar next function) one of the only functions that violate Difference Maker, as some reasons are excluded from consideration.

We can of course also combine both kinds of thresholds: One for reasons to be eligible to contribute, and one for the overall weight of reasons.

Definition 3.4.17 (Double Threshold). *Let k, l be some fixed real numbers.*

$$g_5(R_N) = \{o \in O \mid \sum_{r \in R_o^+ : f(r) > l} f(r) - \sum_{r \in R_o^- : f(r) > l} f(r) \geq k\}$$

This function, unsurprisingly, violates the same principles as either threshold functions (except for Menu Insensitivity, which it satisfies, like g_2 and g_3).

Again, $k = 0$ is an important special case, but we will not discuss this any further here.

The functions so far are all quite simple, which, we think, is appropriate for weighing reasons. But why do they all use sums? We think that this is most in line with the

3. The Kitchen Scale Model Formalized

metaphor of a weighing scale where each reason contributes weight in a linear fashion. However, from a formal standpoint, we are not constrained by these considerations. We could just as well replace the sums by a product in the above definitions or any arbitrary, more complicated function we deem plausible. So, we will include one further simple function that we can compare the above functions to.

Definition 3.4.18 (Product of Reasons).

$$g_6(R_N) = \{o \in O \mid \text{for all } o^* \in O : \frac{\prod_{r \in R_o^+} 1 + f(r)}{\prod_{r \in R_o^-} 1 + f(r)} \geq \frac{\prod_{r \in R_{o^*}^+} 1 + f(r)}{\prod_{r \in R_{o^*}^-} 1 + f(r)}\}$$

This last function is interesting in that using products instead of sums keeps all principles intact save for those that talk about the scaling of weights, and so it violates Relativity – Reasons, Reasons Scaling: Quality, and Reasons Scaling: Quantity.

Principle	g_1	g_2	g_3	g_4	g_5	g_6
All Allowed	yes	no ($k \neq 0$)	yes	yes	no	yes
Anonymity I	yes	yes	yes	yes	yes	yes
Anonymity II	yes	yes	yes	yes	yes	yes
No Effects from Spurious Reasons	yes	yes	yes	yes	yes	yes
No Valence Flips for Reasons	yes	yes	yes	yes	yes	yes
Independence of Reasons	yes	yes	yes	yes	yes	yes
Weak Reason Reciprocity	yes	yes	yes	yes	yes	yes
Strong Reason Reciprocity	yes	yes	yes	yes	yes	yes
Relativity – Reasons	yes	yes	yes	no	no	no
No Dilemmas	yes	no	no	yes	no	yes
Reasons Scaling: Quality	yes	yes	yes	only r^+	only r^+	only r^-
Reasons Scaling: Quantity	yes	yes	yes	only r^-	only r^-	only r^+
Difference Maker	yes	yes	yes	no	no	yes
Reasons Symmetry I	yes	no	no	yes	no	yes
Reasons Symmetry II	yes	no	no	yes	no	yes
Menu Insensitivity	no	yes	yes	no	yes	no

Table 3.2.: Overview of which principles the functions $g_1, \dots, g_6$ satisfy. Proofs are simple, but still included in Appendix A.3.

Let us briefly recapitulate what we have done so far. We have introduced four functions plausibly governing the final weight of reasons and six functions for aggregating the weight of reasons and deriving deontic statuses based on that aggregate weight. We have shown that, in principle, we can exclude some combination of principles altogether. With this, we conclude the principle-based analysis. So far, we have remained rather agnostic about which functions or principles we deem acceptable, and we will continue to be so in the future. Still, we think the simplest function for deriving deontic statuses

g_1 is the most fruitful one for expositional purposes, and so we will continue use it in the future. In the next section, we explore philosophical applications and reply to some potential objections.

3.5. Challenges

In this section, we will address challenges to the model. We will divide them into two (rough) categories of challenges to *weighing* – challenges to the choice of functions f and g once the weight has been put in place and challenges to *weighting* – challenges to the idea that we can make sense of the very idea of giving a specific weight to reasons and other normative considerations at all. We will deal with these in turn.

3.5.1. Challenges to Weighing

First, we will address challenges to weighing. We can distinguish between two distinct types of challenges here. The first questions whether the way we weigh reasons is adequate to actual normative cases out there. In particular, the challenge is to explain, how the structure and functions can account for cases where (i) the joint weight of two reasons is less than simply adding up the weight (in the extreme case, where adding a second reason, doesn't seem to change the weight at all) and (ii) putative cases, where the polarity of a reason changes with context. The second type of challenge has to do with the way we introduced the functions. Are the functions we've introduced not arbitrary? Is there any actual normative difference between (let's say) f_2 and g_4 and f_3 and g_1 ? In order to answer this latter challenge, we will define a sort of consequence relation for Normative Graphs.

Challenges to Aggregation

Let's start with the first type of challenge.³⁶ We will discuss three different versions of this type of challenge, each dealing with the question of how, if at all, we can accurately capture phenomena where the joint weight of two reasons is not just their sum. These three challenges concern, the phenomenon of *reasons overlap*, the phenomenon of *diminishing returns*, and the (supposed) phenomenon of *polarity flips*.

Let us start with the first challenge – *reasons overlap*.³⁷ Reasons overlap describes the

³⁶These challenges are also found in (Tucker, 2023), who discusses them in a similar, but non-formal, setting.

³⁷(Nair, 2021) poses this as his second challenge for any model of reasons. We take the following example

3. The Kitchen Scale Model Formalized

phenomenon where there are two reasons for an action, but (a part of) one reason is contained in another. The easiest way, to make this concrete, is by giving an example: The fact that Adam is in pain is a reason to help him. The fact that Adam is in severe pain is also a reason to help him.

This version, where one reason completely encompasses another, is also the most difficult to get right. Here, arguably, the weight of both reasons combined is not the sum of their weight (or determined by any of the functions we’ve proposed in Section 3.4.2) but instead simply the weight of the reason given by the fact that Adam is in severe pain. How can we deal with this? We need, as (Tucker, forthcominga) suggests, to avoid *double counting* the reason(s).

There are two ways to respond to this challenge. One is to reject it and one is to extend the model. We think both are plausible responses, and we will describe them in detail below.

By rejecting the challenge, we mean claiming that the phenomenon of reasons overlap needs to be addressed *outside* the model. It is a task for *modeling*. This means that the task of accounting for the phenomenon falls on people creating the Normative Graph. A Normative Graph with two reasons, one of which contains the other, would then simply be inadequate and should be remedied by explicitly excluding any reason completely subsumed by another. We think this is a plausible answer. Any model has limitations, and if we understand them, we can account for them when deploying the model. But still, this answer is a little unsatisfying.

The second way to respond to the challenge is to explicitly tackle the phenomenon inside the model. To do this, however, we need to extend it. The main idea, which strikes us as very plausible, is that the reason that is given by Adam’s severe pain is undercutting the reason given by Adam’s pain. But we then need to explain (inside the model) why this is so. We think there is a way to do this if we add some further machinery and extend the model with a logical language instead of the mere symbols we used as propositions until now.

Recall, that we called our set of propositions L , using the same letter usually used for logical languages. Indeed, we could replace these mere symbols by the set of propositions (let’s stick with calling it L) from a logical language $\mathcal{L}$.³⁸ With this change, we can keep all our other definitions intact.

What we are looking for is a constraint on our graphs that ensures that a *larger* reason always undercuts a *smaller* reason. We can model this, like we modeled other constraints

from (Kearns, 2016).

³⁸We have not done so because, until now, adding a logical language on top, was an entirely unnecessary complication.

on the set of *reasonable* graphs, as a constraint on the set $\mathcal{N}$ collecting these graphs. Then, let $\vdash_{\mathcal{L}}$ be the consequence relation of $\mathcal{L}$. We want that given a Normative Graph $N = (\Delta, O)$ that contains $\delta_1 = (p, c, x)$ and $\delta_2 = (p', c, x)$, both of which are reasons for or against the same action, with $p \vdash_{\mathcal{L}} p'$, there is a $\delta_3 = (p, \delta_2, u)$ contained in N as well (such that for a given weighting function w , $w(\delta_3) > 0$).³⁹ So what this constraint says is that the more *specific* reason always wins and undercuts the less specific reason. We think this is exactly what happens in cases of reasons overlap, and that constraint is a natural solution to avoid double counting.

We think either response, rejecting the challenge or extending the model, is acceptable, but even if we want to tackle the phenomenon inside the model, we can. Unfortunately, there is a related challenge that is harder to solve, where there is no overlap of reasons, but the joint weight of reasons is still not just the sum of their weights – the phenomenon of diminishing utility. Consider the following example.

Your colleagues Adam and Benoît teach the same class. Both offer you 10 chocolate croissants each to teach the class next week. This gives you two reasons to teach the class. The first given by the fact that Adam will give you 10 croissants and the second given by the fact that Benoît will give you 10 croissants. These reasons have the same strength. But, since having more and more chocolate croissants has diminishing utility for you (you will become sick of them at some point), the strength of both reasons together will be less than simply double their individual contribution.^{40,41}

What is the problem here? It appears that the reasons do not add up completely, but there clearly is no overlap to explain it. On the face of it, none of our aggregation functions can account for this.

Here, rejecting the challenge does not work, as clearly, both reasons have to be included in the model for it to be useful. Unfortunately, we also have no solution that addresses this in a principled way like before.⁴² But the model can still deal with these cases, if we take care to model the case correctly and if we embrace the principle Inclusive Holism, as suggested by (Tucker, 2023). In cases of diminishing utility, (at least one of the) facts that ground reasons also serves as an attenuator of another reason. The fact that Adam will give you 10 croissants makes the reason given by the fact that Benoît will give you 10 croissants less weighty. And analogously, the fact that Benoît will give you 10 croissants makes the reason given by the fact that Adam will give you 10 croissants less weighty. So, when modeling these cases, we need to include these attenuators explicitly

³⁹In effect, we presuppose the principle Inclusive Holism with this solution. We will discuss this again below.

⁴⁰We follow (Tucker, forthcominga) in describing this class of cases as cases of *diminishing utility*.

⁴¹This is best understood if one considers the function g_1 (see Definition 3.4.13), but the example can be made to work with other aggregating functions as well.

⁴²We think, this is because there is no principled way to solve this. Cases of diminishing utility depend on how much the utility is diminished, and there is no a priori way to know how much that is.

3. The Kitchen Scale Model Formalized

in the Normative Graph to correctly represent normative reality. Unlike in the case of overlap, however, there is no constraint on the set of graphs that guarantees that we have modeled a situation correctly.

Here, our model can accommodate the phenomenon of diminishing returns, but only if we accept a liberal version of what Normative Graphs can look like. We need to accept Inclusive Holism and allow that some facts ground both reasons and modifiers.

The last challenge, dealing with supposed polarity flips, is best explained by a case from (Prakken and Sartor, 1996). The argument for polarity flips runs as follows. The fact that it is hot is a reason against jogging and the fact that it is raining is a reason against jogging, but the fact that it is hot and raining is not only not a reason against jogging, it is a reason *for* jogging because in these conditions jogging is pleasurable. Both reasons together seem to not only lose some of their weight, they seem to switch polarity. Clearly, this is something we don't allow (see principle 3.4.23). If this is a genuine case, we are in trouble.

Luckily, this is dubious. The most common critique is that neither the fact that it is hot nor the fact that it is raining is what is at the normative bottom. They are not *basic reasons*.⁴³ The actual reason giving element is something more primitive, like (dis)pleasure. The reasons would then look something like this. The fact that running is not pleasurable is a reason against running (in a context where it is either (only) hot or raining) and the fact that running is pleasurable is a reason for running (in a context where it is (both) hot and raining). The fact that it is hot, raining, or both is then only causally relevant to that pleasure.⁴⁴ What this means for us is that it simply depends on factors outside the model, which Normative Graph best represents the case. If we want to model a case where it is either hot or rainy, the appropriate graph for this situation contains the reason grounded by the fact that running is not pleasurable, if it is both hot and rainy, the graph would include the reason grounded by the fact that running is pleasurable.

We think that this distinction between basic and derived reasons tracks some of our intuitions rather well. But can we give a proper argument as well? We think so. Assume that we are right and that facts about (dis)pleasure really *do* ground reasons – this seems uncontroversial to us. We might still say that the fact that it is raining (or hot, or both)

⁴³See (Tucker, forthcominga) who argues in a very similar context against heat and rain and see (Berker, 2018) to see an independently motivated use of the distinction between basic and derived reasons.

⁴⁴There is a certain worry here that we are using the notion of basic reasons to obscure rather than illuminate. How do we know what basic reasons are? Do any of the examples we have used so far contain basic or derived reasons? The usual answer (see for example (Tucker, forthcominga) for this exact answer in this exact context) is that this is for first-order ethicists to figure out, and an independent question of how to model these things. However, if we could not readily identify reasons, modifiers and undercutters in everyday situations, the usefulness of the model we present here is diminished. Nevertheless, we think that the distinction between basic and derived reasons helps us reject the *specific* case of heat and rain.

gives rise to *additional* reasons, and we might run into trouble again.

This is a strange position to hold. Assume that (for whatever reason) it would still not be pleasurable to run even if it were hot or rainy (and you have no other reasons, like those related to your fitness, to go for a run). It strikes us as very implausible to say that you have *any* reason whatsoever to go for a run. And even if one can make this plausible, we would need to say that the fact that it wouldn't be pleasurable would *outweigh* or, more likely, *undercut* the reasons given by the facts that it is hot and rainy. This strikes us as an even more absurd position to take, as there doesn't seem to be any normative connection at all between facts about the weather and pleasure. The causal relation doesn't seem to have anything to do with it.

This is not a knock-down argument against cases like heat and rain. In fact, what would be needed is an argument to show that polarity flips cannot ever occur. We are not aware of any such argument, and we will not attempt to give one here. Should we be wrong here, we would have to give up principle 3.4.23 (and maybe 3.4.5 for modifiers as well) and propose new semantics. This presents a serious challenge that nevertheless still can be tackled with some modifications.⁴⁵

So we have seen that we can give a direct answer to two of the first challenges from the first group that relies on accepting Inclusive Holism for both challenges. To solve the problem of reasons overlap we need to accept that a fact that grounds a reason can also ground an undercutter and for the harder problem of *diminishing returns*, we need to accept that a fact that grounds a reason can also serve as a modifier. Our model, on the other hand, cannot easily accommodate polarity flips. Here, we think we are right to deny that this is a genuine phenomenon. Should it exist, however, the model we present would need to be, and can be, reworked.

Are the Functions Arbitrary?

The second group of challenges asks whether the different functions we have proposed might not be completely arbitrary. One can think of this challenge as claiming that whatever function(s) we propose, they will not make a normative difference in the end. Perhaps, there simply is no right function.⁴⁶ The task is then to show not only (i) that the functions we have introduced are different, but (ii) give a principled account of how

⁴⁵One option that we will only sketch here is this: In cases like heat and rain, what is happening is that heat not only attenuates (or undercuts) the reason given by rain but also changes the role of the reason from a reason against to a reason in favour. In this case, we would need to extend our model not only to include a final weight, but also something akin to a *final role*. This is not technically difficult to include, but complicates the model and might itself be controversial, so we prefer to keep the simple model intact.

⁴⁶This particularist (in the style of (Dancy, 2004)) objection, however, need not be the underlying motivation here.

3. The Kitchen Scale Model Formalized

we can come to find out and (iii) give a way to *falsify* certain functions.

There is one obvious answer to this challenge that one can give. We have undertaken a principle-based analysis after all, and clearly, each of the functions satisfies a different set of principles. So, clearly the functions are, at least in that sense, sufficiently different. And if we have independent reason to accept one of the principles, that is reason to reject any function that violates it.

But the principles' content is the weight of reasons and other considerations, and as we have mentioned, in the case of the function governing the final weight of reasons, there is almost nothing to be found in the philosophical literature as to how exactly this function should behave. The situation is slightly better in the case of reasons, where there is some disagreement over how they should be weighed with some rare, concrete proposals for how it is to be done. Still, to fully satisfy a sceptic, what we want is to show that the functions make a concrete difference when it comes to the normative verdicts in some cases, independently of the potentially controversial assignment of weights.

So showing that is what we will do here. We will introduce a kind of consequence relations on cases that allows us to perform casuistic reasoning using Normative Graphs. We then show that this reasoning depends on the choice of the functions, and that they do make a genuine difference here. It also allows us to *reject* functions. If we think that what we should do in some set of cases determines what we should do in another (set of) case(s), but according to the semantics we use the set of cases does not determine what we should do in the other case (or vice versa), then we have good reason to think that the functions are wrong.

So, the next step is to make this casuistry formally precise. The first step is to define what a *case* is.

Definition 3.5.1 (Case). *A case is a tuple (N, A) where $N = (\Delta, O)$ is a Normative Graph and $A \subseteq O$ is a set of actions.*

A case thus represents what actions are morally permissible (A), in a certain situation (N). Note, that a case does not include any information about the specific weights of reasons or other normative considerations.

The next step is to define a consequence relation on cases. The idea is that if we know what the morally correct actions in some cases are, then, logically, we also know what we should do in some other set of cases.

Definition 3.5.2 (Satisfaction).

- Let (N, A) be a case, w a weighting function, and $f \in \{f_1, \dots, f_4\}, g \in \{g_1, \dots, g_6\}$. We say w satisfies (N, A) (given f and g) iff $g(R_N) = A$, given f and w .
- Let $\mathcal{N}$ be a set of cases. We say that w satisfies $\mathcal{N}$ (given f and g) iff w satisfies every case $(N, A) \in \mathcal{N}$ (given f and g).

Definition 3.5.3 (Consequence). Let $\mathcal{N}$ be a set of cases and (N, A) be a single case. Let $f \in \{f_1, \dots, f_4\}$ be a function governing the final weights of normative considerations and $g \in \{g_1, \dots, g_6\}$ be a function governing the deontic statuses. Then:

- We say that a case (N, A) follows from a set of cases $\mathcal{N}$ (given f and g) – written as $\mathcal{N} \vdash_{f,g} (N, A)$ iff every weighting function w that satisfies $\mathcal{N}$ (given f and g) also satisfies (N, A) (given f and g).
- We call a set of cases $\mathcal{N}$ inconsistent (given f and g) iff there is no weighting function w that satisfies $\mathcal{N}$ (given f and g).
- We say that a case (N, A) is consistent with a set of cases $\mathcal{N}$ (given f and g) iff $\mathcal{N} \cup \{(N, A)\}$ is not inconsistent.
- We generalize these definitions to sets of cases in the standard way.

These definitions allow us to compare the functions by their consequence relations. Unfortunately, because we have 4 functions governing the final weight of considerations and 6 functions for determining the deontic statuses of actions based on those final weights, this gives us a total of $4 * 6 = 24$ consequence relations. In order to show that they are all different, we would need to show the pairwise difference of $23 + 22 + \dots + 1 = 276$ combinations. This is too long, even for the appendix of this thesis, and so, even though we strongly endorse the following statement, we must leave it as a conjecture here.

Conjecture 3.5.1. For every combination f, g and f', g' such that $f, f' \in \{f_1, \dots, f_4\}$ and $g, g' \in \{g_1, \dots, g_6\}$ if it is not the case that both $f = f'$ and $g = g'$, then $\vdash_{f,g}$ is not identical to $\vdash_{f',g'}$

We can prove a slightly less ambitious statement, and at least show that *one* of the function combinations is different from all the others; we will start from the beginning and prove the statement for f_1 and g_1 .

Proposition 3.5.2. The consequence relation $\vdash_{f_1, g_1}$ is not identical to any consequence relation $\vdash_{f, g}$ with $f \in \{f_1, \dots, f_4\}$ and $g \in \{g_1, \dots, g_6\}$ where not both $f = f_1$ and $g = g_1$.

3. The Kitchen Scale Model Formalized

The proof for this proposition is not complicated, but quite long and can be found in Appendix A.3.

So let us recapitulate what it means for functions to be sufficiently different. We said that we can compare them on two levels. First, we can compare them on the level of principles. Here, we make plausible assumptions about how default weights determine the final weight of considerations, and how we weigh reasons. We can also compare them based on which *normative inferences* they allow one to make. This is what we did when we introduced the consequence relation on cases. In either case, we see that the functions are not identical (even though, for the second part, we stopped short of proving that every possible consequence relation is different from every other). Nevertheless, we think that these two ways to think about functions allow us to answer the challenge. The functions are different, and we have two levels on which we can *falsify* functions. First, if we can show that a principle that we consider important is violated by a function, and second, if the normative inferences the function allows us to make are implausible.

With this, we conclude the part on challenges to *weighing*. We have seen that the model gives us the tools to show why the joint weight of two reasons does not necessarily need to equal the sum of their weight, and we have seen how we can decide (at least to some extent) between the functions we want to use to calculate the final weights and the functions we want to use to weigh the reasons.

3.5.2. Challenges to Weighting

The second group of challenges comprises challenges to *weighting* reasons (and other considerations). We will address these challenges next. There are two main challenges raised in this group, the first by (Nair, 2021) and the second by (Drai, 2018). We will present each challenge in more detail below, but first let's see what they're about. Nair's challenge consists in making sure, that we can give a satisfying explanation of what the weight of a reason even *means*. Drai's challenge is more radical: she claims that there simply is no way to come to a determination of weights at all. The answer to both of these challenges will lie in showing that we can *reverse engineer* the weighting function.⁴⁷ The insights of which will in turn allow us, in Chapter 4, to apply the model as a basis for machine ethics.

Nair's Challenge

The challenge of (Nair, 2021) is to “[provide] a suitable representation of the strength of reasons”(Nair, 2021, p. 7). Nair, in his paper, claims that, so far, *reductive approaches*

⁴⁷We take the term reverse engineering for this approach from (Tucker, 2023) and (Tucker, forthcominga).

– approaches that reduce the notion of a reason to some other concept (prominently, in Nair’s paper, decision theoretic concepts) can answer this challenge, while *non-reductive approaches* – approaches that take the notion of a reason as primary, so far, cannot. Nair proposes that one way to overcome this challenge is to give a set of qualitative axioms, much like the ones for probability theory or decision theory.

We claim that, at least in part, our model gives an answer without using probabilistic tools, but in order to do this, we need to make the challenge a little more explicit.

Our answer will, unsurprisingly, depend on the semantics we put on reasons. That is, each function g that takes us from the set of reasons to the deontic statuses of actions will give a slightly different answer as to what reasons mean. For the sake of simplicity, we will use the most straightforward function as an illustration here: The *Sum of Reasons* function g_1 (see Definition 3.4.13). Given g_1 , can we show that we have provided a “suitable representation of the strength of reasons”? It is not enough to simply point to the model and say it uses weights to determine the deontic statuses of actions in such and such a way, and then show that it can be fruitfully applied in ethics. This is not what Nair has in mind. Rather, what we must, and indeed will show, is that the weights of *specific* reasons are not just something we can willy-nilly posit, but something that is metaphysically necessary (at least in a range of cases) if g_1 turns out to be true.

What tools can a non-reductionist avail themselves of to rise to this challenge? First, we believe (and Nair appears to agree) that we can claim that reasons, modifiers and undercutters are fair game as long as they don’t include specific weights from the outset. That is, we can take the notion of a Normative Graph as given.

Second, perhaps slightly more controversial, we can make use of (one of the) semantics we have introduced as long as we have made them independently plausible. That is, for example, we can use one of the functions f governing the final weight of considerations and one of the functions g determining the deontic statuses of actions based on these final weights.⁴⁸ From these two elements, we then need to show that we can make sense of *specific weights* in *specific* situations that have *specific* normative meaning. What we cannot do is simply put some numbers on reasons that fit the deontic verdicts we want to see in some case. That is, we need some independent arguments for the choice of a *weighting* function w . We will provide these after discussing a related challenge.

⁴⁸Nair might claim that we have so far not done enough to make these functions plausible. When discussing (Sher, 2019), he seems to suggest that for a model of reasons based on confirmation theory, he wants a set of representation principles that fix the weights of reasons. But while we encourage finding a more axiomatic representation of the functions we propose, these functions govern the *weighing* of reasons, not the *weighting*. And so we believe that to rise to Nair’s challenge, having a plausible semantics is enough. What we do need to be precise about is what weights we assign once the semantics is in place.

3. The Kitchen Scale Model Formalized

Drai's Challenge

(Drai, 2018) argues that “reasons have no weights”. A part of her argument makes the claim that any account of the final weights of reasons based on default weights of reasons and context factors is explanatorily empty. Next, we will present Draï’s argument as a challenge to our account. In section 3.5.2, we will respond to Draï’s challenge.

Drai attacks the idea that there can be a “moderate holist account of weights”. Unfortunately, Draï does not give a definition of what makes an account moderate, but we can piece together a faithful definition.⁴⁹

The easiest way to set up what a moderate account of weights is in Draï’s terminology is to contrast it with a *radical* holist account of weights. A radical account says that the weight of reasons depends on the entire context they are located in. It is radical insofar as it says that there is nothing we can say about the weights of reasons until we have seen the entire context.

A moderate holist account then denies that. While the context still has some influence, we can know the context-dependent weight by just looking at parts of the context. Arguably, our *constrained holism* is moderate in Draï’s sense, as we can calculate the final weight of normative considerations from the leaves up and do not need to consider *all* normative considerations in a graph.

Drai makes two related but slightly different arguments to support her claim that reasons have no weight. The first one, and the one she is explicit about, is essentially epistemic.

[W]e cannot have [a moderate account] for one’s capacity to recognize weights of reasons (Drai, 2018, p. 11)

The second argument that implicitly shines through in Draï’s paper is essentially metaphysical.⁵⁰ A moderate account of the weights of reasons is uninformative. In a moderate holist account, the weights of reasons are arbitrary and serve no explanatory purpose. Each case will have a set of weights on the reasons that allow you to derive the correct deontic statuses, but there is no general theory behind it.

The main premises for Draï’s arguments in both challenges are the same. Draï calls looking at a reason isolated from any (other) context the *isolation test*. She then establishes

⁴⁹Drai uses an analogy with semantics to motivate the term. Specifically, between “radical and moderate accounts of meaning”. Where on the radical account the meaning of the elements of a sentence depend on the truth value of the entire sentence, while on the moderate account they do not.

⁵⁰We take it that is an argument that Draï makes in the paper, even though she does not announce it. Should that not be an accurate representation of Draï, it is still an important, related challenge to consider.

(correctly) that:

The isolation-test cannot help us to understand [the notion of a default weight], since a reason which is isolated from all other reasons has only a direction towards a certain action; there is nothing to add that might correspond to weight. (Drai, 2018, p. 13)

Here, the *direction towards a certain action* describes the polarity of a reason. So far, this is uncontroversial. the next step in her argument is likewise something we can support:

[T]he notion of a weight of a reason is a comparative notion; it is based on comparisons between reasons (A is a stronger or weightier reason than B). (Drai, 2018, p. 14)

This corresponds to our principle 3.4.11. Not all functions satisfy it, but we can take this claim on board. In the next section, we will show that even using function g_1 – a function that does satisfy this principle – Drai’s challenge does not succeed.

Since looking at reasons in isolation does not give us any weights, Drai claims that, in order to determine the weights of reasons in one situation, we have to look at counterfactual situations – different cases where the reason is the same, but the context is different.

The only way to answer [the question of what the weight of an isolated reason A is] is by looking at the counterfactual situation, where we put A next to another object and compare between them. This gives us the weight of A in the counterfactual situation. However, if Holism is true, the weight of A in the counterfactual situation tells us nothing about the weight of A in the original one. (Drai, 2018, p. 14)

This, in Drai’s view, means that

If the weight of an object depends on the context of weighing, and we weigh objects by comparing them to other objects, the question: “What is the weight of A when there are no other objects to compare it to?” is meaningless. (Drai, 2018, p. 14)

This is the part we take issue with (and indeed we will show that it is wrong). But before we do that, let’s bring Drai’s challenge into sharp relief. The challenge is this: How can we, from a set of situations, determine the default weight of reasons (and other normative considerations)? Again, we can read this as an epistemic challenge: How can

3. The Kitchen Scale Model Formalized

we come to know the (default) weight of reasons from a set of situations? We can also give it a metaphysical reading: Can we show that default weights arise necessarily (and not just arbitrarily) out of a set of situations?

We will address this challenge in both its readings in the next section.

Reverse Engineering Weights

Here, then, we rise to meet the challenges by Nair and Draï. We will do this by showing that (at least in some cases) we can reverse engineer *precise* weights of reasons and other normative considerations. This allows us to answer Nair's challenge and reject Draï's claim that weights in our moderate holist account are meaningless.

We will give a simple example of how one can come to know *precise* values of reasons and modifiers. Most cases we encounter in usual circumstances will neither be as simple nor allow for such a precise determination of weights. So consider the following a proof of concept that it is in principle possible to determine the weights of (some) reasons and modifiers precisely, even though for others this might not always be possible. To keep the example simple, we will keep using the simplest function for the weighing of reasons g_1 (see Definition 3.4.13) as our aggregation function.⁵¹

In the end, what we will show is the following propositions.

Proposition 3.5.3. *Given g_1 and some $f \in \{f_1, \dots, f_4\}$, for every $q \in \mathbb{Q}$ and every two reasons r, r' , there is a finite set of cases $\mathcal{N}$, such that for every weighting w function satisfying $\mathcal{N}$, $w(r) = q * w(r')$.*

That is, in some situations, we can determine the weight of reasons precisely (up to scaling with some common factor). Similarly, for the second proposition:

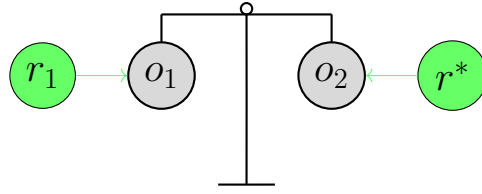
Proposition 3.5.4. *Given g_1 and some $f \in \{f_2, \dots, f_4\}$ for every $q \in \mathbb{Q}$, for every (non-undercut) modifier δ , there is a finite set of cases $\mathcal{N}$, such that for every weighting function w satisfying $\mathcal{N}$, $w(\delta) = q$. For $f = f_1$, we can determine the weight to be $w(\delta) = q * w(r)$, where r is some reason.*

That is, for modifiers, in some cases, we can even figure out the *precise* weight.

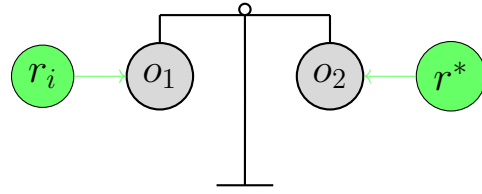
⁵¹We can adapt the example for the other functions as well, but it would become more complicated. We will mention on which principles this particular example relies below, but slight adaptations are possible where they do not hold. The only thing we cannot guarantee is that we can always give a concrete value to all reasons, but only rank them on a ratio scale with unknown origin.

Usually, in this thesis, we banish the proofs to the appendix, but the way we prove these statements is both relatively straightforward and instructive, and so we will go over them in some detail here.

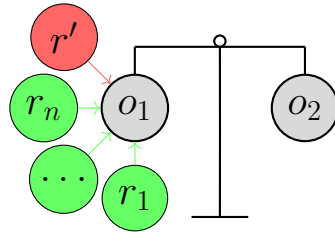
We start with a very simple example of just reasons and no modifiers and a case with two options $O = \{o_1, o_2\}$. We assume (to make this example simple and tractable) that Principle 3.4.11 holds; that is, the weights of reasons are relative and can be arbitrarily scaled up or down. So, without loss of generality, we say that $w(r_1) = 1$ where $r_1 = (p_1, o_1, r^+)$ is a reason for o_1 . Additionally, we have a reason $r^* = (p^*, o_2, r^+)$ for o_2 with (so far) unknown weight. Let us further assume that in this Normative Graph $(\{r_1, r^*\}, O)$ all options are allowed. Then (by principle 3.4.32 and 3.4.26) we obtain that the reason r^* must have the same strength and so $w(r^*) = 1$.



Next, we repeat this process $n - 1$ times. For every reason $r_i = (p_i, o_1, r^+)$, for $i \leq 2 \leq n$, such that in the graph containing just r_i and r^* both options are permitted. By the same argument as before, this means that $w(r_i) = w(r_1) = w(r^*) = 1$.



Now consider a further reason against o_1 , $r' = (p', o_1, r^-)$, such that we know that in the graph $(\{r_1, \dots, r_n, r'\}, O)$ still all options are permitted. Hence $w(r_1) + \dots + w(r_n) = n * w(r_1) = n = w(r')$. That is, n times the weight of r_1 .



So we can give a series of cases that guarantees that a reason has n times the weight of another reason, but then, to show that the weight of two reasons r and r' is such that $w(r) = \frac{a}{b}w(r')$, all we need to do is add further cases (generated just like the ones above) where we have a reasons of weight $\frac{1}{b}$ that are equal to the weight of r . And so we have a proof for the first proposition.

3. The Kitchen Scale Model Formalized

Clearly, this is a contrived example, but with cases like these it is possible to determine the weight of a reason to be some specific rational number (up to scalar multiplication of all reasons involved).⁵² So, what we have established so far is that the weights of reasons (for some choice of functions f and g) are uniquely determined up to scalar multiplication in some cases. That means, that (at least for those cases and those weights) the weight of the reasons is not arbitrary. We will discuss the implications of this further below, but for now, we need to quickly address modifiers.

In fact, we can construct similar cases that determine the weight of modifiers *precisely*. As a proof of concept, we will use the simplest case (and function) we can get away with.

Assume we have again two options $O = \{o_1, o_2\}$ and one reason in favour of each option that we have already determined the weights of (perhaps with cases like the one above) $r_1 = (p_1, o_1, r^+)$ and $r_2 = (p_2, o_2, r^+)$, such that $2 * w(r_1) = w(r_2)$. Furthermore, we have an amplifier of r_1 , namely $\delta = (p_3, r_1, a^+)$, such that in the Normative Graph $(\{r_1, r_2, \delta\}, O)$ all options are again permissible. We then know that $f(r_1) = w(r_3) = 2 * w(r_1)$. This gives us that the $w(\delta) = 1$, given $f \in \{f_2, f_3, f_4\}$ and $w(\delta) = w(r_1)$ for $f = f_1$. In the latter case, as long as we fix the weight of a reason to some number, we also fix the number for f_1 , for all the other possible choices of f , no matter the weight of the reasons, the weight of the modifier is already fully determined, just by the relative strengths of reasons in this case. It's easy to generalize this proof idea to arbitrary $q \in \mathbb{Q}$. The idea can also be straightforwardly applied to determine the weight of modifiers of modifiers, deeper down in the graph.⁵³

Clearly, we will not be in a position to determine the weights of reasons and modifiers just by looking at cases all the time. Occasionally, we are left with a complicated set of (in)equalities that constrain what values the considerations take. But what the above example shows is that it is *in principle* possible to reverse engineer the weights of reasons and other considerations in such a way that we obtain a *precise* weight for all of them.

Answering the Challenges to Weighting

We believe that the possibility of reverse engineering the weights of reasons and modifiers allows us to give a unified answer to Nair's and Draï's challenges. Let us start with Nair's challenge. The challenge Nair posed was that non-reductive theories of reasons need to be able to give a "suitable representation of the strength of reasons". We said that the non-reductionist can avail herself of the structure of normative cases (that is,

⁵²We conjecture that it is not possible (with finite cases) to determine the weight of a reason that is in $\mathbb{R} \setminus \mathbb{Q}$ for any non-trivial function satisfying the principle F-Restriction.

⁵³As long as they do not occur in a chain of considerations that includes undercutters; then, all bets are off. Luckily, in such a case, the exact weight of modifiers is not relevant for the deontic statuses we can derive from such a graph.

Normative Graphs). Arguably, she can also avail herself of a function f and g as, as long as the choice is motivated, as there is nothing reductive about how reasons and other normative considerations interact. Given these ingredients, we need to come up with a good story about what the weights of reasons *actually mean*. And we think, we can give this answer quite concisely: at least in some (sets of) cases, the weight of reasons (and other considerations) is the exact thing that allows us to derive what we ought to do in these cases. That is, given a set of cases, only if we posit the weight of reasons (and other considerations) to be such and such, can we derive what we ought to do in these cases. The set of cases pseudo-axiomatically fixes these weights.⁵⁴

We think that for Draï's challenge, too, we can provide an answer along these lines. Draï claims that there is no way to determine the default weights of reasons given a moderate holist account. But we have seen that this is simply false. At least in some cases, we can give as precise an answer to the question of what the weights are as it is possible to give (that is, we can precisely determine their relative strength). The ingredients we needed for this are all things that Draï herself allows us to use. A set of cases (without any weights), the morally correct actions in these cases and a theory about weighing reasons (including, in our case, a theory of how the final weights we use for weighing are determined). With this, we can show two things: first, given these ingredients, in some (set of) cases, the weights of (some) reasons are metaphysically necessary. In fact, we have shown that for every reason, there is some set of cases that necessitate it having a certain weight. Secondly, we have also given a story about how we can come to know the weights of reasons. If we have independent grounds to believe that we should perform some actions in some cases, then, if we are lucky, these cases uniquely determine the weights of reasons (and modifiers); at least up to scalar multiplication.

So let us sum up what was discussed in this section. We claimed that the model can answer the two challenges surrounding *weighing*. We answered the first challenge on how to model cases where the weights of two reasons do not simply add up, by postulating that sometimes the facts that ground reasons can also serve as attenuators or undercutters of reasons. We answered the second challenge, claiming that the different functions we propose do not make a true difference, by showing that they licence different normative inferences for cases.

For the second set of challenges, due to Draï and Nair, we showed that we can reverse engineer the weights of reasons (at least in some cases). This allowed us to show three things. First, we could show that the weights of reasons, in some cases, are the exact thing that allow us to determine the set of morally correct actions, which unambiguously fixes their meaning. Second, we showed that, in some cases, the weights of reasons are a necessary by-product of the semantics we propose, and third, we have given a concrete

⁵⁴Note that this does not entail that we are explaining the weights of reasons in terms of oughts. We are saying that the meaning of the weights of reasons is how they uniquely determine the relationship between the structure and the oughts in a set of cases. We do not need to accept any metaphysical or explanatory priority of oughts over reasons to make this claim.

3. The Kitchen Scale Model Formalized

way how one can come to learn those weights by reasoning through cases. Hence, we think, our model has allowed us to produce novel philosophical results as well. We have seen that we can give a non-reductive account of the weights of reasons and rebut Drai's claim that a moderate holist account of reasons and their weight is impossible.

3.6. Justifying & Requiring Reasons

In this section, we will look more closely at extending the model to allow for the distinction between justifying and requiring reasons. Here, in the non-formal case, we have (Tucker, 2021, forthcominga) as guidance.

We will first show how we can include justifying and requiring reasons, and then analyze in more detail how they behave. Then, we will show how we can model the three puzzling cases at the heart of Chapter 2 in our model and highlight some differences between the two approaches.

In setting up the formal model without any distinction between justifying and requiring reasons, we went through four steps. We set up the structure of normative considerations, their default weights, their final weights, and a way to get from the final weights to deontic statuses of actions. In order to make room for justifying and requiring reasons, we can skip some of these steps as they, arguably, remain unchanged. The first idea one might have is that we only need to amend the notion of a Normative Graph slightly. This would allow us to change the way deontic statuses of actions are derived only a bit. Indeed, even while we will later amend this definition, it is a good first step.

Definition 3.6.1 (Dual-Role Reasons (Preliminary)). *A (dual-role) reason is a tuple (p, c, x) , where $p, c \in L$, $x \in \{r^+, r^-, j^+, j^-\}$. Like before r^+, r^-, j^+, j^- are special symbols representing, respectively, whether a reason is a requiring reason for, a requiring reason against, a justifying reason for, or a justifying reason against an action.*

We define the functions $Premise(\cdot)$, $Conclusion(\cdot)$, and $Role(\cdot)$ as before, to return the first, second, and third element of the tuple.

Note that like in our model in Chapter 2, we *extended* the notation to include justifying reasons and that the notation for requiring reasons mirrors that for the model containing only a single type of reason. This, like in the previous chapter, is not an accident and the correspondence between requiring and single-type reasons will be easy to spot, once we have a formal model in place.

The notion of *justifying reasons against* might seem a little mysterious at first. After all, it is the job of requiring reasons to make an action permissible and talking about

justifying reasons against might seem to imply that they (like requiring reasons against) make actions *less* permissible. We will discuss what this notion means later. For now, we will simply take them as a formal artifact.

The definition for normative considerations can remain unchanged (if we interpret the set of reasons as above). The notion of a Normative Graph can then also be kept the same. Analogously, the weighting function will assign weights to both justifying and requiring reasons (and all other normative considerations).

We will assume that modifiers and undercutters, too, behave the same for both kinds of reasons and the way the final weight of justifying and requiring reasons (and all other normative considerations) can be determined by one of the functions $f \in \{f_1, \dots, f_4\}$.

For the actual *weighing* of reasons, we will focus on one of the functions (g_1) and extend it to include the distinction between justifying and requiring reasons. We focus on a single function here for two reasons. First, the ideas we present carry over easily to the other functions and second, the function is very close to the one defended in (Tucker, forthcominga) and it will be interesting to highlight the differences.

As we have discussed in more detail in Chapter 2, justifying reasons, as we understand them, metaphorically push an action towards permissibility, but they do so without making its alternatives less permissible. Any scale that models this kind of weighing adequately will look utterly broken. So let's break a scale.

Let us briefly recall g_1

Definition 3.6.2 (Sum of Reasons).

$$g_1(R_N) = \{o \in O \mid \text{for all } o^* \in O : \sum_{r \in R_o^+} f(r) - \sum_{r \in R_o^-} f(r) \geq \sum_{r \in R_{o^*}^+} f(r) - \sum_{r \in R_{o^*}^-} f(r)\}$$

In order to extend this function to justifying and requiring reasons, we need to first define the sets of justifying and requiring reasons that we are going to put on the scale. This definition is quite similar to the one, we defined for the single-type model in Definition 3.4.10.

Definition 3.6.3 (Justifying and Requiring Reasons for an Option). *Let $N = (\Delta, O)$ be a Normative Graph (based on dual-role reasons).*

- *Let $J_o^+(N)$ – or just J_o^+ – be the set of all justifying reasons for o , that is, $J_o^+ = \{j \in \Delta : \text{Conclusion}(j) = o \in O \text{ and Role}(j) = j^+\}$*

3. The Kitchen Scale Model Formalized

- Let $J_o^-(N)$ – or just J_o^- – be the set of all justifying reasons against o , that is, $J_o^- = \{j \in \Delta : \text{Conclusion}(j) = o \in O \text{ and } \text{Role}(j) = j^-\}$
- Let $R_o^+(N)$ – or just R_o^+ – be the set of all requiring reasons for o , that is, $R_o^+ = \{r^* \in \Delta : \text{Conclusion}(r^*) = o \in O \text{ and } \text{Role}(r^*) = r^+\}$
- Let $R_o^-(N)$ – or just R_o^- – be the set of all requiring reasons against o , that is, $R_o^- = \{r^* \in \Delta : \text{Conclusion}(r^*) = o \in O \text{ and } \text{Role}(r^*) = r^-\}$
- As before, the set R_N represents the set of all reasons in N , hence $R_N = \bigcup_{o \in O} (R_o^+ \cup R_o^- \cup J_o^+ \cup J_o^-)$.

To ease us into it, we first ignore all justifying reasons against actions. A good first attempt to capture how justifying and requiring reasons relate (in the style of g_1) is the following (preliminary) function g_1^* .

Definition 3.6.4 (Sum of Reasons* (Preliminary I)).

$$\begin{aligned}
 g_1^*(R_N) &= \{o \in O \mid \text{for all } o^* \in O : \\
 &\sum_{r \in R_o^+} f(r) - \sum_{r \in R_o^-} f(r) + \sum_{r \in J_o^+} f(r) \\
 &\geq \\
 &\sum_{r \in R_{o^*}^+} f(r) - \sum_{r \in R_{o^*}^-} f(r)\}
 \end{aligned}$$

Here, the (joint) weight of justifying reasons pushes an action towards permissibility by lending it its weight, but it doesn't impact the other actions at all. We can think of this as extending what we will call the *permissibility interval* of an action. For some action o , let $\mathbf{R}_o$ be the joint weight of the requiring reasons for o , that is $\mathbf{R}_o = \sum_{r \in R_o^+} f(r) - \sum_{r \in R_o^-} f(r)$ and $\mathbf{J}_o^+$ the joint weight of the justifying reasons for o , that is $\mathbf{J}_o^+ = \sum_{j \in J_o^+} f(j)$, then the permissibility interval is $PE(o) = [\mathbf{R}_o, \mathbf{R}_o + \mathbf{J}_o^+]$. That is, the minimum in the joint weight of the requiring reasons and the maximum is the joint weight of the requiring reasons together with the justifying reasons. We can then determine the deontic status of an action by looking at this interval. Specifically, if, in a choice between two actions, the *maximum* of this interval is higher than the *minimum* of the other action, the action is permissible. As an illustration, see Figure 3.3.

Thinking about justifying reasons this way, we think, is a good starting point. But this function has some serious problems. To bring them into the light, we will discuss justifying reasons *against* next.

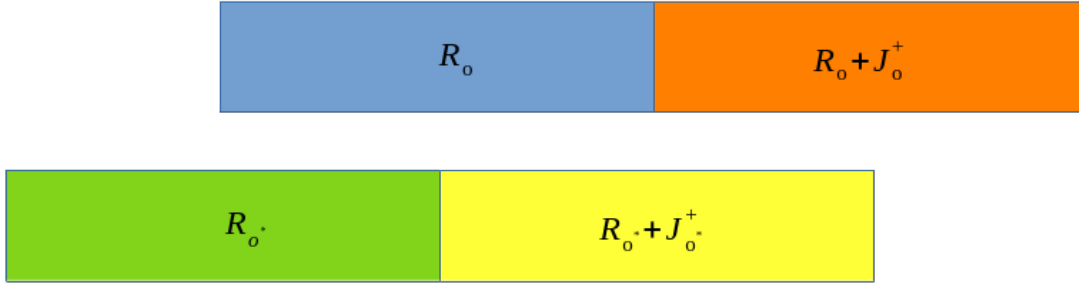


Figure 3.3.: A permissibility interval, showing two (equally) permissible actions o and o^*

3.6.1. Justifying Reasons Against

So what can we say about justifying reasons against? First, we will make them plausible conceptually by discussing cases where they occur. Second, we will discuss how they affect the deontic status of actions.

So what about justifying reasons against? Do they exist? Can we give an example? How do they affect the deontic statuses of actions?

Justifying reasons are often used to explain (supposed) phenomena of permissible partiality.⁵⁵ Take for example:

Death at Sea Your child is drowning to your left. To your right, two children that you have never met are drowning. You can either save your own child or both of the other children.

Here, proponents of permissible partiality claim that the fact that the child is your child *justifies* you saving it and thus justifies letting the two children drown. So we find, somewhere, a justifying reason in this example.⁵⁶

If we accept that partiality is a good fit for justifying reasons, we need to allow for justifying reasons against. The reason is that partiality swings both ways. Consider the following example.

⁵⁵For a defence of this view (in not quite the terminology used in this thesis), see (Stroud, 2010), (Scheffler, 2010), and (Lord, 2016).

⁵⁶What exactly the premise of the reason is a tough question and will probably require further analysis: Luckily, we don't need to model this case faithfully. (Tucker, 2022) argues that what happens here is a modifier adding *justifying weight* to a reason. We prefer to remain agnostic, as long as there is some kind of justifying reason (or any reason with justifying weight) to be found.

3. The Kitchen Scale Model Formalized

Death at the Office Two equally racist coworkers are choking on their food at the office. They're quite racist, but not quite racist enough to deserve to die. You know two facts about these two coworkers. The first coworker has a cat, and the cat will have to find a new home if that coworker dies, while no one depends on the second co-worker. However, the second coworker's racism targets *your* ethnic group, while the first coworker's targets another. You can only save one of them.

Here, if there is anything to the permissible partiality position at all, you are allowed to let the coworker whose racism targets *you* die and save the other, even though this means that a cat needs a new home (otherwise, the cat would be the decisive factor in whom to save, we assume). The reason is a justifying reason *against* saving the cat-loving coworker. She hates *your* ethnic group in particular, and that means her life is forfeit!

So our argument goes that if permissive partiality exists and can be explained in terms of justifying reasons, then justifying reasons also come in a negative form. But how do we weigh them? The example above provides a clue. A justifying reason against an action makes the *other* options more permissible *without* making the action they are reasons against any less permissible. So once we understand this, we can propose (the unfortunately still preliminary) following definition

Definition 3.6.5 (Sum of Reasons* (preliminary II)).

$$\begin{aligned}
 g_1^*(R_N) &= \{o \in O \mid \text{for all } o^* \in O : \\
 &\sum_{r \in R_o^+} f(r) - \sum_{r \in R_o^-} f(r) + \sum_{r \in J_o^+} f(r) \\
 &\geq \\
 &\sum_{r \in R_{o^*}^+} f(r) - \sum_{r \in R_{o^*}^-} f(r) - \sum_{r \in J_{o^*}^-} f(r)\}
 \end{aligned}$$

So here, the justifying reasons against an action have no (direct) impact on the permissibility of an action, but they do have an impact on the permissibility of all *other* actions. In particular, this means that we can find an important distinction when it comes to the *canceling out* of reasons. For single-type reasons (and hence for requiring reasons, as they behave identically in the absence of justifying reasons according to our preliminary definition) the principle Strong Reason Reciprocity holds, saying that if we remove a reason for and against an action that have the same (final) weight, the deontic status of an action does not change. The same does *not* hold for justifying reasons as justifying reasons for and against *target* different actions. This unassuming feature, will turn out to be a major problem.

We can best make sense of this if we, again, think of g_1^* as spanning a permissibility interval. Let (additionally to the definitions before) $\mathbf{J}_o^- = \sum_{j \in J_o^-} f(j)$. Then,

the permissibility interval of an action o the interval $PE(o) = [\mathbf{R}_o - \mathbf{J}_o^-, \mathbf{R}_o + \mathbf{J}_o^+]$. Then in a choice between the two actions o and o^* , o is permissible if, and only if $\max(PE(o)) \geq \min(PE(o^*))$.

Regardless of how we think about justifying reasons, we see a parallel in this model to the model of Chapter 2. Requiring reasons take over the same role as what we just called *reasons* in the model without any distinction between requiring and justifying reasons, while justifying reasons do something new.

3.6.2. The Puzzles Revisited

We have said that the definitions so far are still preliminary. We can see why (and how) they need to be amended by revisiting the puzzling cases surrounding supererogation we introduced in Chapter 2. Let's check if our model can accommodate them.

Let us revisit the first case:

Save One or Two. Adam and Benoît are in mortal danger. You can easily and without cost to yourself save one. You can also save both, but that would involve serious harm to you: you would lose your legs.

We can model this case as a Normative Graph quite easily. We can choose between four actions that we collect in O_1 . First we can simply do nothing, which we will represent as o_0 as we save no one. Second, we can save Adam or Benoît but not both, which we represent as o_A and o_B respectively. Third, we can also save both Adam and Benoît, which we represent as o_{AB} .

There are more reasons than in our formalization in Chapter 2. This will be the first clue we have. The first two reasons we need are requiring reasons to save Adam and Benoît. We represent them as $r_1 = (AdamsLife, o_A, r^+)$ and $r_2 = (BenoitsLife, o_B, r^+)$ respectively. They are clearly *requiring reasons* for saving their lives. However, we clearly want them to be reasons for saving *both* as well. But, so far, we need to include these facts as independent reasons. And so we include $r_3 = (AdamsLife, o_{AB}, r^+)$ and $r_4 = (BenoitsLife, o_{AB}, r^+)$ as well. The last reason, on the other hand, is a justifying reason to keep our legs. We think the most intuitive way to model this reason is as a justifying reason against saving both Adam or Benoît, as it appears that the reason makes it permissible not to perform this action.⁵⁷ And so we represent this reason as $j_1 = (LegsImportant, o_{AB}, j^-)$. To capture the intuition that we are permitted not to save both Adam and Benoît if it costs us our legs, we set $w(r_1) = w(r_2) = w(r_3) =$

⁵⁷We discuss whether it makes a difference to model cases like these with justifying reasons for or against when we talk about the next case in footnote 62.

3. The Kitchen Scale Model Formalized

$w(r_4) > 0$ and $w(r_1) + w(r_2) \leq w(j_1)$.⁵⁸

This gives us the Normative Graph $N_1 = (\{r_1, r_2, r_3, r_4, j_1\}, O_1)$. We can calculate the permissibility interval for all actions. For o_0 it is simply $PE(o_0) = [0, 0]$ as no reason speaks in favour or against it. For o_A and o_B , we have the interval $PE(o_A) = PE(o_B) = [w(r_1), w(r_1)]$, but for o_{AB} we get the interval $PE(o_{AB}) = [w(r_3) + w(r_4) - w(j_1), w(r_3) + w(r_4)]$. It is easy to see that o_0 is not permitted according to our definition as $0 = \max(PE(o_0)) < \min(PE(o_A)) = w(r_1) > 0$. But both o_A and o_B are permitted as $\max(PE(o_A)) = \max(PE(o_B)) = w(r_1)$ is larger than both $\min(PE(o_0)) = 0$ and $\min(PE(o_{AB})) = w(r_3) + w(r_4) - w(j_1) < 0$. We can see that o_{AB} is permitted as well, as $\max(PE(o_{AB})) > \min(PE(o_A)) = \min(PE(o_B)) > \min(PE(o_0))$.

This means that our preliminary definition is able to model the *Save One or Two* case perfectly. We used this case in Chapter 2 to generate the first paradox surrounding supererogation, which we called the *Classic Paradox of Supererogation*. It asks how it can be that an action that seems best is not also obligatory. The answer given in chapter two was that this is because justifying reasons make it permissible not to do it. This is, of course, also the answer here, but we can do something better: We can actually rank actions in terms of their goodness.⁵⁹

Definition 3.6.6 (Goodness of Actions). *Let $N = (\Delta, O)$ be a Normative Graph, f a function governing the final weight of considerations. Then an action $o \in O$ is at least as good as $o^* \in O$ iff $\sum_{r \in R_o^+} f(r) - \sum_{r \in R_o^-} f(r) \geq \sum_{r \in R_{o^*}^+} f(r) - \sum_{r \in R_{o^*}^-} f(r)$.⁶⁰*

This gives us a (total) order on actions in a Normative Graph. It also allows us to model cases of supererogatory holes explicitly (as we will see below).⁶¹ We can then say that the action one *Ought* to take, is (one of) the best action(s). This is well in line with what we have called the *standard account* of supererogation, except that it allows for the existence of supererogatory holes quite naturally.

⁵⁸This makes it easier to reuse the same reasons for the next case as the intuition behind these cases is that you don't have to save either Adam or Benoît if saving a single person would cost you your legs either.

⁵⁹What we are about to do should not be confused with trying to analyze the notion of goodness in terms of reasons in the style of buckpassing accounts of value (see (Rabinowicz and Rønnow-Rasmussen, 2004)). We are not offering an analysis of goodness, we merely claim that there is some correspondence between goodness and reasons.

⁶⁰This is a definition of goodness leaning towards g_1 as the weighing function, but it should be clear how a similar definition would look that is more similar to some other $g \in \{g_2 \dots g_6\}$.

⁶¹(Tucker, forthcominga) suggests something along these lines, but ultimately rejects it. He likes to hold fast to the principle that an action is permissible if it is better than another permissible action. To be able to do this, he introduces the concept of *commending weight* of reasons, which (potentially) can be determined by a kind of lexicographic ordering of reasons based on their requiring weight and their deontic status. We think this is a strange, ad-hoc, and entirely unnecessary manoeuvre on Tucker's part.

So this model allows us to distinguish between the three modalities that we introduced in Chapter 2 as well. You *ought* to do an action iff it is the best action. You are *required* to perform any of the permissible actions (and so a single action is required iff it is the sole permissible action). And you are permitted to do an action iff it is the set $g_1^*(R_N)$.

These definitions will carry over to the final version of our definition of g_1^* . But let us first see if we can accommodate the other two puzzling cases. The next such case is the *All or Nothing* case.

All or Nothing. Adam and Benoît are trapped in a collapsing building, and you will lose your legs whether you save one or both of them.

We can model the Normative Graph that corresponds to this case similarly to the one above. Let $N_2 = (\Delta_2, O_2)$ where O_2 contains the same four actions as O_1 above. And again, the requiring reasons in Δ_2 are the same four reasons $r_1, \dots, r_4$ (with the same weights) as before. What is different in this scenario is the justifying reason. Here, in order to represent the scenario using justifying reasons against, we need *three* reasons.⁶² These are $j_2 = (LegsImportant, o_A, j^-)$, $j_3 = (LegsImportant, o_B, j^-)$, $j_4 = (LegsImportant, o_{AB}, j^-)$. Like before, we assume $w(j_2) = w(j_3) = w(j_4)$ and this weight is larger than the joint (requiring) weight for saving either Adam or Benoît.

The problem with this formalization is that we don't get the desired results. Now, even saving just one person is permissible. The reason lies in how we think about the permissibility intervals. In the (binary) choice between o_A and o_{AB} clearly, o_{AB} should win out. The reasons against each action are the same, but there is an extra reason (saving Benoît) that speaks in favour of o_{AB} . But in the way we do it until now, we compare the minimum of the permissibility interval for o_{AB} with the maximum of the permissibility interval for o_A . That is, we compare $w(r_1) + w(r_2) - w(j_4)$ with $w(r_1)$. But since the weight of the justifying reason is strong o_A wins out and becomes permissible. This is not how it should be.

The phenomenon we encounter can be described as *undercounting* justifying reasons. The same justifying reason gets counted in favour of just one action, namely o_A , but not for o_{AB} . But, arguably, if the same reason speaks in favour of or against any two actions, when we compare these two actions, the reason should lend its weight to both actions equally. But with our formalization, this is not what happens. As a reason can only be a reason for a single action, we have to split them into multiple parts. This means we have no way to tell when a reason for two different action is ultimately *the same reason*. Note that this is a problem that uniquely affects justifying reasons. In the case of requiring reasons (or just single-type reasons) we see the same phenomenon of

⁶²There is the alternative to represent this justifying reason as a single reason *for* walking away. If we chose to go the other route, we will not run into the problem below *now*, however for the next paradox, the Intransitivity Paradox, the problem would reappear. So, here, we will keep with the (we think) natural formulation to motivate the final version of g_1^* .

3. The Kitchen Scale Model Formalized

splitting up reasons (as can be seen in N_1 and N_2) but requiring reasons have a property that justifying reasons do not. Every function we proposed in Section 3.4.2, satisfied the principle Strong Reason Reciprocity stating that a reason for and against with the same strength *cancel out*. In the case of the function g_1 , we could additionally model reasons for (or against) other actions as reasons against (or for) its alternatives (through principle Reasons Symmetry I and Reasons Symmetry II). These two facts together mean that if we encounter a requiring reason for some action and a requiring reason for some other action, the weight of these two reasons will simply disappear if we compare these two actions. In the case of justifying reasons, on the other hand, the weight persists.

What we propose is taking the insight from above to solve the problem. It appears that the reason for saving Adam is the same reason, whether it speaks in favour of saving just Adam or in favour of saving both Adam and Benoît. So it appears that sometimes the same reason can speak in favour of more than one action.⁶³ And making this explicit, allows us to solve the problem of undercounting as well.

So let's amend the notion of a reason to include this fact (giving us the final definition of dual-role reasons).

Definition 3.6.7 (Dual-Role Reasons). *A reason is a tuple (p, A, x) , where, like before, $p \in L$ is the premise of the reason, $x \in \{r^+, r^-, j^+, j^-\}$ its role. The difference is that now $A \subseteq L$ is a finite set of actions.*

We can keep the other definitions intact, considerations are still defined using the concept of reasons, premises return a single element of L , conclusions are now subsets of L . We again want to make sure that in every Normative Graph all conclusions of the reasons are indeed possible actions, but no major changes need to be done.⁶⁴ Likewise, the functions $f_1, \dots, f_4$ can be kept unchanged.

Let's see what that means for the All or Nothing case: our new Normative Graph $N^* = (\Delta^*, O)$ is much more tractable. We only need three reasons in this graph: one (requiring) reason for saving Adam, $r_A = (AdamsLive, \{o_A, o_{AB}\}, r^+)$ and one for saving Benoît, $r_B = (BenoitsLife, \{o_B, o_{AB}\}, r^+)$ and a single (justifying) reason for not doing either, $j_L = (LegsImportant, \{o_A, o_B, o_{AB}\}, j^-)$. Again, we assume that $w(r_A) = w(r_B)$ and that $w(j_L)$ is larger than the sum of these two requiring reasons.

⁶³We believe that this is a different phenomenon than the same *premise* grounding a reason for two different actions. Take as an example the following: You don't like being hot, and it is summer. Then, the fact that it is sunny is a reason for you to eat ice cream and to stay inside. But it is not the same reason in both cases. If it is impossible to do both, you might even compare the weights of both reasons to find out what you ought to do.

⁶⁴We also keep the weighting function the same. Potentially, we would like to assign a different weight for each possible action that the reason speaks in favour of, but we leave this complication aside for now.

Of course, to define our weighing function, we need to make some minor changes to Definition 3.6.3 where we defined what the reasons for an option are. For example: the set of requiring reasons speaking in favour of an action o , in some Normative Graph (Δ, O) , R_o^+ is now the set $\{r^* \in \Delta : o \in \text{Conclusion}(r^*) \text{ and } \text{Role}(r^*) = r^+\}$. But none of the changes require substantial rewriting.

What this allows us is to throw out reasons appearing twice for multiple actions, and thus allows us to eliminate undercounting.⁶⁵ Hence, we can define the final definition of our new function g_1^* taking into account justifying and requiring reasons.

Definition 3.6.8 (Sum of Reasons*).

$$\begin{aligned}
 g_1^*(R_N) &= \{o \in O \mid \text{for all } o^* \in O : \\
 &\sum_{r \in R_o^+ \setminus R_{o^*}^+} f(r) - \sum_{r \in R_o^- \setminus R_{o^*}^-} f(r) + \sum_{r \in J_o^+ \setminus J_{o^*}^+} f(r) \\
 &\geq \\
 &\sum_{r \in R_{o^*}^+ \setminus R_o^+} f(r) - \sum_{r \in R_{o^*}^- \setminus R_o^-} f(r) - \sum_{r \in J_{o^*}^- \setminus J_o^-} f(r)\}
 \end{aligned}$$

And, it is easy to check, that (as long as all the weights of reasons are positive) we get exactly the desired result. o_{AB} is permitted as it is not outweighed by any action, o_0 is permitted because of the justifying reason against performing any other action (and thus losing your legs), but o_A (and o_B) is not permitted, as:

$$\begin{aligned}
 &\sum_{r \in R_{o_A}^+ \setminus R_{o_{AB}}^+} f(r) - \sum_{r \in R_{o_A}^- \setminus R_{o_{AB}}^-} f(r) + \sum_{r \in J_{o_A}^+ \setminus J_{o_{AB}}^+} f(r) = 0 \\
 &< \\
 &\sum_{r \in R_{o_{AB}}^+ \setminus R_{o_A}^+} f(r) - \sum_{r \in R_{o_{AB}}^- \setminus R_{o_A}^-} f(r) - \sum_{r \in J_{o_{AB}}^- \setminus J_{o_A}^-} f(r) = w(r_B) > 0
 \end{aligned}$$

The best action is saving both (and thus, what you ought to do), it is permissible to do nothing and to save both, but forbidden to save just one person. To see this definition in action, let us consider the final puzzling case involving supererogation.

Life or Promise You've promised a friend to meet her for dinner. En route to the restaurant, you come across a collapsing building, with Adam trapped inside. Saving him will cost you your legs.

⁶⁵This idea bears some similarity to the comparativist move in (Tucker, forthcominga). Here, Tucker makes the controversial claim that what a reason is, depends on the pairwise comparison with another action. We will discuss this below in Section 3.7.

3. The Kitchen Scale Model Formalized

We can model this similarly with a Normative Graph $N_3 = (\Delta_3, O_3)$. Where $O' = \{o_W, o_P, o_A\}$ where o_W stands for the action of walking away, o_P stands for the action of keeping your promise (and going to the dinner) and o_A stands for the action of saving Adam.

Δ is composed of the reason to keep your promise $r_P = (Promise, \{o_P\}, r^+)$ and the reason to save Adam $r_A = (AdamsLife, \{o_A\}, r^+)$. We have a justifying reason for you to keep your legs, which justifies not taking o_A and so we represent this reason as $j_L = (LegsImportant, \{o_A\}, j^-)$. Again, to keep the example plausible, we assume $w(r_P) < w(r_A) \leq w(j_L)$.

We can easily confirm that this gives us the desired results. It is both permitted to keep our promise and to save Adam, and the best thing we could do is saving Adam. Our model gives us the right results for this case as well.⁶⁶

So let's quickly recap the journey so far. In this section, we discussed how to include a distinction between justifying and requiring reasons. We claimed that if one accepts that partiality requires justifying reasons, then we need to allow for justifying reasons against actions. Unfortunately, simply adding justifying and requiring reasons in the style of single-type reasons didn't work as sometimes, we would count the same reason only for one action when comparing two actions. To make room for this, we had to change the definition of what a reason is, making it possible (in some cases) to cancel out justifying reasons like we do for requiring reasons. This in turn allowed us to model the cases generating the paradoxes of supererogation from Chapter 2 more naturally, as well as resolve them in a unified manner.

3.7. Related Work

In this section, we will compare the model we presented to other work. As, as far as we know, this is the first formal treatment of the kitchen scale model of weighing reasons, we will mostly compare different aspects of the model to related work. In particular, we will compare the way we calculate the final weight of considerations to related work in formal (weighted) argumentation. We will also compare the way we derive oughts from the (final) weight of reasons with the (non-formal) account of (Tucker, 2021, forthcominga), looking more closely at the way Tucker treats justifying and requiring reasons. Finally,

⁶⁶We have not talked about the paradoxes these puzzles generate here. In fact, it would be easy, but somewhat tedious, to define a conditional modality that conditionalizes on *available* actions, that is subsets of O . Given a Normative Graph $N = (\Delta, O)$, we would then say that an action is permitted given $O' \subseteq O$ iff it is permitted in the graph (Δ, O') . This definition would allow us to restate the paradoxes in our model here, and it is straightforward to show that the paradoxes get resolved in the same way as we discussed in Chapter 2. The kitchen scale model as we have formalized it here implies that we should give up the principles as well.

we will briefly compare the formal kitchen scale model to the decision theoretic account of (Sher, 2019). The relationship of formal models in general to the kitchen scale models was already discussed at the beginning of this chapter, as well as in Chapter 1 and 2.

3.7.1. Weighted Argumentation

The way we calculate the final weight of considerations is most closely matched by work on weighted argumentation (see (Amgoud and Ben-Naim, 2016; Amgoud et al., 2017; Amgoud and Ben-Naim, 2018; Amgoud et al., 2022)). Amgoud et al. define a *bipolar* argumentation framework (A, R^+, R^-) where the set A is a set of *arguments*, or roughly, statements put forth in deliberation or while arguing for a point and the relation R^+ is a relation that models how different arguments support each other. So if an argument a stands in relation R^+ to some argument b , we read this as argument a supporting argument b . Similarly, R^- is an *attack* relation. If an argument c stands in relation R^- to some argument d , argument c attacks argument d , meaning, very roughly, that putting forward c undermines the acceptability of d . The presence of these two relations (support and attack) are what makes this a *bipolar* framework.

What Amgoud et al. intend to model is a quantitative score that represents the “overall strength or overall acceptability” of an argument (see (Amgoud et al., 2017)).⁶⁷ Similar to our model’s considerations, arguments come with a default acceptability score – a real number between 0 and 1, where 0 stands for absolute rejection and 1 stands for absolute acceptance and an argument can change their acceptability score depending on the presence of other arguments, resulting in a *final* acceptability score that remains between 0 and 1.⁶⁸

We can easily map our Normative Graphs to what we might call a *tripolar* weighted argumentation framework. We use the standard (single-type) definition of reasons here.

Definition 3.7.1 (Weighted Argumentation Framework of a Normative Graph). *Let $N = (\Delta, O)$ be a Normative Graph. We call the tuple (A, R^u, R^+, R^-) its corresponding (tripolar) argumentation framework, where A is just Δ , and two elements δ and δ^* stand in the relation R^u iff $\text{Conclusion}(\delta) = \delta^*$ and $\text{Role}(\delta) = u$, they stand in the relation R^+ iff $\text{Conclusion}(\delta) = \delta^*$ and $\text{Role}(\delta) = a^+$, and they stand in the relation R^- iff $\text{Conclusion}(\delta) = \delta^*$ and $\text{Role}(\delta) = a^-$.*

We thus map a Normative Graph to a (weighted) argumentation framework with three

⁶⁷Amgoud et al. are open to what exactly this entails. They cite *certainty degree of [an argument’s] premises, votes, importance, the degree of value promoted, trustworthiness of its source* as possible interpretations. (see (Amgoud et al., 2017)).

⁶⁸We paraphrase the terminology and notation slightly, to make the similarities to ours more transparent.

3. The Kitchen Scale Model Formalized

relations. One is similar to the support relation of Amgoud et al. (our amplifiers) and two are more similar to their attack relation (our attenuators and undercutters).

There are many differences, but also significant similarities, between Amgoud et al. and our approach. We will discuss these next.

The first difference is, of course, found in the field of application. We aim to model normative practical reasons, while Amgoud et al. try to model something more epistemic. For us, the module governing the final weights of reasons is complemented by a module for weighing reasons. There is nothing like it in Amgoud et al.'s framework. We think that our model has a relatively precise way to interpret the semantics: as a formalized version of the kitchen scale model. We have seen that what the weight of a reason means can be given a quite precise answer, albeit one that depends on which function is used to calculate the final weights and to weigh the reasons. We find it more difficult to pinpoint what exactly an acceptability score of, say, 0.3 means. We believe that our semantics stands on more solid ground.

Mathematically speaking, however, Amgoud et al.'s model is more general as it does not impose finiteness or acyclicity restrictions on the framework.

The biggest similarity between the two approaches, however, can be found in the principle-based way functions are compared. Amgoud et al. propose a plethora of principles to compare different functions (just like we do). Indeed, quite a few of the principles we propose can be translated into principles found in (Amgoud et al., 2022) and vice versa. This in turn means that some constraints and principles are independent of the particular area, practical or epistemic, the framework/model is applied to.

3.7.2. Tucker's Dual-Scale Model

The closest and most comprehensive non-formal account related to our model is due to (Tucker, forthcominga). In this book-length defence of the kitchen scale model, Tucker, like us, uses a holist account to defend it against criticism. A (non-formal) version of many of the responses we have given in Section 3.5 can also be found in the book. There are, however, significant differences as well, and we will highlight the most important ones here.

First, Tucker endorses contrastivism about reasons, that is, the thesis that the weight (and perhaps polarity) of a reason for some action is always relative to some other action(s) one compares it to. We have steered clear of incorporating this.

Second, and most importantly, Tucker defends a specific version of how justifying and requiring reasons are weighed: his *Dual Scale* Model (see also (Tucker, 2021)). Comparing this model of weighing reasons to ours is what we will spend the remainder of this

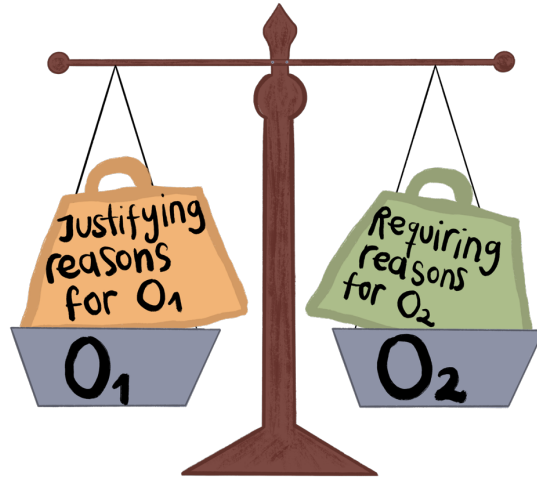


Figure 3.4.: Tucker's Permission Scale

section on.

Reasons, on Tucker's account, come with two types of weights: justifying and requiring. This is a slight departure from the way we added the two types of reasons to our model, treating them as essentially distinct ontological entities.

These two weights then serve different functions (that will sound similar to the one they serve in our model). Justifying weight pushes an action towards permissibility, requiring weight pushes it towards a requirement. We will illustrate his account by only allowing a choice between two actions.

Tucker's main claim is that there is one scale that tells us if an action is permissible – the permission scale – and another if an action is required – the commitment scale. Assume we are to decide between two actions o_1 and o_2 . Then, on both scales, the left pan represents o_1 and the right pan represents o_2 . To figure out whether o_1 is permitted, we look at the permission scale, where the joint weight of the *justifying reasons for o_1* pushes the left pan down and the joint weight of the *requiring reasons for o_2* pushes the right pan down.⁶⁹ If the left pan is lower (or at least equally low) than the right pan, the action is permitted. See Figure 3.4.

Similarly, to figure out if o_1 is required, we look at the commitment scale. Here, the joint weight of the *requiring reasons for o_1* push down the left pan, while the joint weight of the *justifying reasons for o_2* push down the right pan. Again, if the left pan is lower (or

⁶⁹We ignore reasons against for now; they behave symmetrically in Tucker's account.

3. The Kitchen Scale Model Formalized

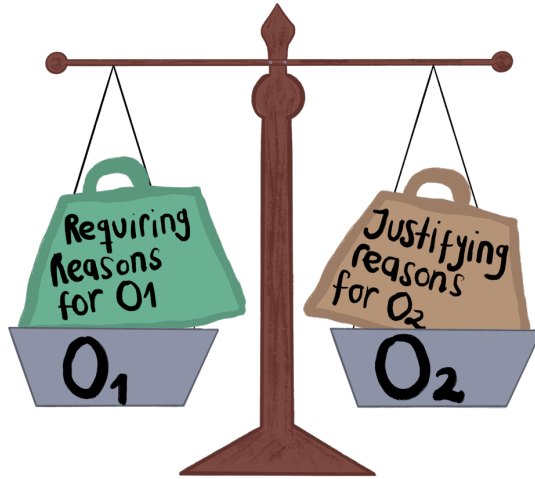


Figure 3.5.: Tucker's Commitment Scale

at least equally low), then o_1 is required; otherwise it is not. See Figure 3.5.

This dual-scale model is more flexible than ours. It is, however, not clear if much is gained by this flexibility. The dual-scale model allows (for example) that an action is required but not permitted (if it's requiring weight relatively high, but its permitting weight is relatively low). In particular, it is possible for genuine dilemmas, where neither action is permissible, to arise on this model. Tucker seems to think this is an advantage of his model, as the claim that no dilemmas exist can now be modeled as the substantial assumption that the justifying weight of reasons is always at least as large as the requiring weight.

Let's compare this to the way we (in the form of the function g_1^*) model this. First, dilemmas are impossible by the definition of the weighing function. If normativity allows for genuine practical dilemmas, then the function is inadequate.⁷⁰ But the existence of such dilemmas is, to say the least, controversial. Furthermore, we actually think it is quite plausible that whether dilemmas are possible is decided on the level of weighing reasons and not on the level of the individual reasons and their weight.

We also think that our model has some advantages over Tucker that make it more intuitive in another dimension. Namely, the way we treat *purely justifying* and *purely requiring* reasons. In our case (at least for g_1^*), requiring reasons take the role of the single-type reasons in the earlier model. What we do, in effect, is *adding* a second type of reason on top. We think this fits well with the arguments often made in favour of the

⁷⁰Should that be the case, we have bigger problems than how best to represent reasons.

existence of justifying reasons. They simply have been overlooked, as many cases just do not contain any such reasons. Requiring reasons behave exactly as expected when we put them on a pan. It is justifying reasons that behave strange in that they only add weight in one direction. Whether one agrees with this or not, we take it as a point in favour of our model that we only have one scale. It is just a little broken.

The last point we need to address is that of Tucker’s comparativism. Comparativism is a variety of the contrastivism that we have already mentioned. It says that what counts as a reason for an action depends on some action we compare it with. As an example, Tucker claims, if you have the option of saving one person or saving the same person and another one, when comparing these two actions, the reason for saving one person disappears. Tucker makes this move to be able to accommodate the All or Nothing scenario. We similarly had to exclude some reasons from popping up in some cases. Are we committed to comparativism of this sort too then?

To solve the All or Nothing puzzle, we did indeed exclude some reasons from adding their weight when we compare the relative weight of (two) certain actions. In our case, we had to include this because otherwise the lack of symmetry (or canceling out) of justifying reasons led to a kind of undercounting. We solved this by excluding identical reasons for two different actions from making a difference when comparing the two actions. This indeed is, or can be interpreted as, a kind of comparativism, but it is a very weak one, as we only exclude the exact *same* reasons from consideration. All other reasons behave normally.

So our model, in contrast to Tucker’s, only needs one scale to account for justifying and requiring reasons. We think it treats (purely) justifying reasons in a more intuitive manner and only presupposes a rather weak form of comparativism.

3.7.3. Sher’s Decision Theoretic Model

In the last part of this section, we will briefly compare our model to decision theoretic models in the style of (Sher, 2019). We have already introduced these models in Section 1.5, so we will only say a few words of recap here.

Sher uses a probabilistic background and models reasons for an action as propositions that raise the expected value of some action conditional on the proposition obtaining. This allows him to model how context can change the weight of reasons and how the weight of two reasons adds up easily, purely by using decision theoretic tools.⁷¹ Sher’s model has some undeniable advantages over ours. It makes use of a well-understood framework (decision theory) and does not need to make any additional assumptions

⁷¹Later in the paper, Sher also discusses how one can go from the weight of reasons to a probability distribution and utility function; but for present purposes this exposition is enough.

3. *The Kitchen Scale Model Formalized*

other than that this is a good way to think about practical normativity. It also allows him to model cases of polarity shifts (like the heat & rain case) quite naturally, but, as we have seen, whether these exist is dubious.

What Sher does less well, is providing a model in line with most of the (non-formal) philosophical literature. What makes a reason a reason, a modifier a modifier, and an undercutter an undercutter, cannot be answered except by pointing to an increase or decrease in conditional utility. But, that at least is frequently claimed in the literature, there is a fundamental difference between a fact speaking in favour of some action (a reason) or some fact increasing or decreasing the normative salience of some fact (a modifier). Our model in turn handles this task effortlessly (as it was designed to do) by treating these elements as entirely distinct and explaining their relation through the graph structure of a Normative Graph.

We also think that we have an additional advantage over Sher, which we have seen in Section 3.5. Our model can give a non-reductive account of the weights of reasons. Reasons, at least in some cases, have a very particular meaning in our model, and furthermore, we can actually come to learn that weight through well-considered normative judgments. If we follow the arguments of (Nair, 2021), this is (so far) lacking in Sher's account.

So let us recapitulate. The way we calculate the final weights of considerations is very similar to the weighted argumentation frameworks proposed by Amgoud et al. Still, both the meaning and the exact way to calculate the final weight differ quite a bit. Tucker, on the other hand, is in some respects quite similar in how we determine the deontic statuses of actions, particularly when including justifying reasons. Still, where Tucker uses the somewhat complicated model of two scales, we can get away with a single (albeit broken) one. This allows us to not have to posit that every reason comes with both a requiring and justifying weight and thus, we think, presents a more intuitive model of how reasons interact. Lastly, while Sher's decision theoretic model can rely on an established theoretical framework, we think our model is both more intuitive and in line with the (non-formal) literature. Additionally, our model is attractive for non-reductionists (and reasons-firsters) as it allows one to have a coherent notion of weights as well as a plausible story of how we could possibly come to learn them.

3.8. Conclusion

To conclude this chapter, let us briefly recap what we have done in this chapter and then discuss potential future work.

To summarize this chapter in one sentence: We have proposed what we take to be the first

formal model of reasons faithful to the model used by large swaths of the philosophical literature on reasons: the kitchen scale model.

We have said that for this model to make any sense it has to be a holist model: the weight of reasons must be able to change depending on the context they are situated in. We have further claimed that the most plausible version of holism is *constrained holism*, where reasons (and other normative considerations) come with a default weight that can only be changed by a select few context factors: modifiers (attenuators and amplifiers) and undercutters, resulting in a context-dependent *final weight*. We have modeled this structurally using *Normative Graphs*.

In order to determine the best way to assign final weights to reasons (and other normative considerations) we have undertaken a *principle-based* analysis and ultimately proposed four functions that we classified based on these principles. We then did the same for functions that take us from the final weight of reasons to the deontic status of actions. Here, too, we proposed and classified six such functions.

With the model fully in place, we applied it to some philosophical puzzles surrounding weight and weighing, and in the process defended it against criticism. The first set of puzzles dealt with *weighing* reasons. We showed that the model is flexible enough to make sense of (at least some) cases where the joint weight of two reasons is not just given by some kind of *additivity*. We also showed that there is a discernible difference between the functions we proposed by introducing a kind of consequence relation on *cases* – Normative Graphs (without any weights) where we know which actions are morally correct.

The second set of puzzles dealt with the problem of the *weight* of reasons and ways to assign such weight – *weighting* reasons. This took the form of two challenges by Nair and Draï, which we claimed can be answered through *reverse engineering*. We proved that (at least in principle) the weights of reasons can be inferred from a set of cases and (at least in some cases) can be interpreted as just the thing that allows for some deontic status in some situation to hold. In doing this, we have also given an epistemic explanation of how we can come to learn the weights of reasons.

Subsequently, we discussed how we can extend the model to incorporate the distinction between justifying and requiring reasons (the topic of Chapter 2). To include this distinction, we claimed that in order to solve the puzzling cases in Section 3.6, we need to reformulate the notion of a reason. We did this in order to eliminate justifying reasons that get undercounted when we compare two actions which they both justify. We argued that we need some kind of canceling of justifying reasons that is different from the way requiring reasons (or single-type reasons) behave. We argued that this constitutes at best a weak form of comparativism.

Lastly, we compared the model we presented to related work in formal argumentation,

3. The Kitchen Scale Model Formalized

due to Amgoud et al., to Tucker’s dual-scale model of weighing reasons and to Sher’s decision theoretic model of reasons.

So, now that the model is set up, what further things can we do with it? We think the model is quite flexible and lends itself to a variety of applications in philosophy and machine ethics, of which we will discuss a few below.

The first two areas of future work deal with the principle-based analysis. So far, we have mostly stayed in the first two stages of principle-based analysis: analysing different semantics for the model. We believe there is more to be done. First, it might well be possible to give qualitative axioms that guarantee that the function weighing the reasons is broadly additive. The challenge, here, is to find principles that are independently motivated and do not simply encode the desired result. The second part concerns further impossibility results. We believe that we have just scratched the surface here, and a host of philosophical incompatibilities between different ways to look at reasons can be found if they are encoded as principles.

Relatedly, we think that giving a full-blown principle-based analysis for justifying and requiring reasons would prove enlightening as well. As we have seen, to get the correct result for at least some cases we need to be able to *cancel out* reasons (or, alternatively, establish a kind of symmetry between justifying reasons for and against). We believe that even the task of creating a set of principles to divide up possible views on justifying and requiring reasons will be helpful to distinguish views on the topic in the literature and carve up the space of possibilities.

Additionally, we believe it might be fruitful to extend the model to a *contrastivist* model of reasons, where reasons are only a reason for some action relative to another action or set of actions. Proponents of contrastivism claim that it offers a specific explanatory advantage over non-contrastivism (see most importantly (Snedegar, 2017)). In the context of dual-role views of reasons, this is defended by (Tucker, forthcominga) who claims that non-contrastivist theories do not offer enough resources to model some cases faithfully. We find these claims dubious, and we think a faithfully formalized version of a contrastivist theory of reasons will allow us to prove this formally. On the one hand, we think that often a contrastivist model can be redescribed as a non-contrastivist one and vice versa. Should this turn out to be false, we believe this too would allow us to prove quite general results about the limits of expressibility in both frameworks.

One further field of application is applying the model outside traditional philosophy, in the field of machine ethics. If we are correct, in that what we have proposed in this chapter is indeed a faithful formalization of the kitchen scale model and if the kitchen scale model does indeed say something profound about normativity, then, surely, it is beneficial to use it as a basis for a moral machine – an artificial agent that can represent and reason about normative information and ultimately reach morally correct decisions. Luckily, this future work will not have to wait much longer. In the next chapter, we will

propose a hybrid model making use of machine learning to learn normative information on top of our model.

4. Machine Ethics from First Principles

4.1. Introduction

We concluded the previous chapter suggesting that it might be fruitful to apply the formalized kitchen scale model to the field of machine ethics. This is what we will do in this chapter: build a framework for machine ethics from first principles. The framework we will propose is a *hybrid* framework – it will comprise both symbolically represented normative information and use machine learning tools.

The research questions of this chapter are thus: First, what does a machine ethics framework look like that is philosophically informed by the kitchen scale model? Second, is such a framework performant and scalable, and how well can it predict the morally correct outcomes in a variety of cases?

In order to answer these questions, we will, in Section 4.2, briefly go over the state of the art in machine ethics and introduce the three types of machine ethics frameworks: *top-down*, *bottom-up* and *hybrid*. In Section 4.3, we will introduce the main idea behind our framework: using a set of cases to infer the weights of normative considerations and then using those weights to predict the correct actions in novel situations. In Section 4.4, we implement this idea using genetic algorithms. We discuss genetic algorithms and the details of the implementation there. In Section 4.5, we discuss how we can, in principle, evaluate how good the predictions of our framework are. We introduce a testing apparatus based on *synthetic data*. In Section 4.6, we run four experiments to test and empirically evaluate the framework. In Section 4.7, we then compare our approach to related work, focusing specifically on (Freedman et al., 2020), (Anderson and Anderson, 2018), and (Canavotto and Horty, 2022) as the most closely related prior work. In Section 4.8, we briefly discuss possible future work and conclude.

4.2. Machine Ethics: State of the Art

Machine Ethics has a lofty goal: creating an artificial moral agent that can ‘act’ and ‘decide’ in ethically acceptable ways. The only reliable way to reach that goal is if the

4. Machine Ethics from First Principles

artificial moral agent has two separate but related capabilities: It must be able to (i) *acquire* and (ii) *store and represent normative information* in a machine-implementable way (see (Canavotto and Horty, 2022)). The model we have presented so far allows us to store and represent normative information in two parts: a graph structure based on normative considerations and weights on these considerations. This chapter of the thesis will give a partial answer as to how an artificial moral agent can acquire that normative information.

Before we delve deeper into that topic, though, we will give a rough overview of the field of machine ethics as it stands today. The approaches to machine ethics can be divided into three families (see (Allen et al., 2005)), each giving a slightly different answer as to what it means to *represent* and *learn* normative information. The first family comprises top-down approaches. Top-down approaches encode normative information in a symbolic formalism, usually a logical system, specifying a set of rules. Bottom-up approaches use machine learning (or related quasi-statistical tools) to learn normative information using a diverse set of training data. Hybrid approaches aim to overcome the pitfalls of top-down and bottom-up approaches by combining the strengths of both of them. Usually, this means that normative information is stored symbolically, but learned via the methods of machine learning. Next, we will discuss each of these families in more detail.

Top-Down approaches try to tackle the problem of machine ethics by focusing on the *representation* side of the problem. For approaches falling within this family, normative information is represented *symbolically* often using a logical language that can explicitly represent statements about permissions, obligations, or requirements.

In this family, we find, for example, (Ganascia, 2007). Ganascia attempts to model different ethical systems (including an Aristotelian and Kantian ethics) about whether it is justified to lie in certain circumstances. (Ganascia, 2007) make use of Answer Set Programming to tackle these tasks, by modeling different ethical systems as defeasible rules. This in turn allows one to *compute* the outcome of a given input according to some specified ethical system.¹

Other more recent examples include the ethical governor of (Arkin, 2009). Arkin’s ethical governor is a proposed module for autonomous weapons whose purpose is to prevent unethical (or unlawful) use of lethal force. Whether lethal force is ethical is determined on the basis of three factors, first, the perception and background knowledge of the system, i.e. interpreted sensor data and a generalist database, a (long-term) set of ethical constraints about the rules of war and a (short term) mission-specific set of rules that can be added by the remote operator of the weapon. From these three elements, a deontic logic module then determines whether an action requiring lethal force is ethical.

¹This attempt echoes the attempts of (Bench-Capon et al., 1987; Sergot et al., 1986) to represent legal information using logic programming.

Only if it is, will the weapon autonomously perform this action.

The ethical layer architecture of (Bremner et al., 2019; Fischer and Dennis, 2023) similarly tries to implement an *ethical layer* in the control structure of an autonomous system that, roughly, is situated between the planning and execution stage of the system. The authors model an autonomous system using a Belief-Desire-Intention (BDI) logic that is implemented in Python. The ethical layer is then made up of multiple submodules, which we will not discuss in detail here.² For present purposes, it is enough to state that the ethical constraints of the system come as logical rules that are explicitly stated in the logical language.

Lastly, there is the LogiKEy framework by (Benzmüller et al., 2020). We can think of it as a both more concrete and more general implementation of systems like Arkin’s ethical governor. LogiKEy’s aim is to be a platform for different deontic logic systems. This is achieved by embedding them semantically in Higher-Order Logic (HOL) and implemented using the interactive proof assistant Isabelle (see (Paulson, 1986) for the original implementation of Isabelle). This means that different deontic logics (together with different knowledge bases, if they are logically structured) can be used and combined for applications. Due to the integration of various automated theorem provers in Isabelle, this allows automated checking whether, for example, some action is prohibited or permitted given the logical system and relevant facts. Benzmüller et al. argue that this can be used in a modular fashion on top of some autonomous system much like Arkin proposes for his governor module.

There are clear advantages to top-down approaches. First, the normative information in such a system has a clear meaning that can be interpreted and used by both humans and machines. Second, the output of a system and the way the output is reached is, at least in principle, transparent to a user of such a system. There are clear disadvantages to top-down approaches as well. As normative information has to be encoded explicitly, this has to be done manually and for the system to be usable, every possible situation the system could find itself in has to be planned for. This means that in practice such a system is either limited in the area it can be used in (and not too many factors can determine the outcome) or it becomes increasingly unreliable as factors not thought of by the designers play a role in determining the correct decision such a system should reach.

The second family comprises bottom-up approaches. Bottom-up approaches try to tackle the problem of machine ethics by focusing on the *learning* side of the problem. Here, we find frameworks that attempt to learn or synthesize normative information from a large set of training data using machine learning or other quasi-statistical tools. Unlike in the

²These submodules are a simulation module that simulates the (outcomes of) actions and returns an ethical evaluation, an ethical decision module and a planning module, that together determine if further action is required (and if so, refer back to the simulation module) and otherwise release a set of actions for execution in the system.

4. Machine Ethics from First Principles

case of top-down approaches, the normative information cannot readily be represented symbolically with clear meaning or interpreted by humans. How the learning (and representation) of normative information proceeds largely depends on the tools used.

In the family of bottom-up approaches, we find various ways of using reinforcement learning to learn ethical behaviour (see, for example, (Abel et al., 2016; Balakrishnan et al., 2019; Russell et al., 2015; Wu and Lin, 2018)). Reinforcement learning, very roughly, works as follows. An agent can perform a set of actions, and will do so with a certain probability. After each action, a reward is assigned to the state the agent finds herself in. The goal of a reinforcement algorithm is then to find an optimal set of probabilities to perform a certain action. That is, ideally, with each step, the algorithm learns how to obtain a bigger expected reward given the probabilities with which it performs the actions available to her. To be clear, the exact manner how ethical information is learned is different for different frameworks based on reinforcement learning, but the central idea stays the same. Normative information is encoded in the probability distributions with which an agent performs some action.

A second example in the family of bottom-up approaches was made possible through the advent of large language models. The most prominent (and arguably capable) framework using large language models is Delphi (see (Jiang et al., 2021)). The Delphi framework is trained on select sources of (supposedly) ethical decisions in a wide range of cases, sourced from the social aggregator and forum website *Reddit*. The decisions to take in these cases were then determined by clickworkers through Amazon’s Mechanical Turk. In total, the training material comprises 1.7 million cases in natural language. Through this large training dataset Delphi demonstrates impressive generalization capabilities in the face of novel situations.

The advantages and disadvantages of bottom-up approaches are the opposite of those for top-down approaches. The first advantage is that no (complex) manual encoding is necessary. Second, even complex information that would not have been thought of by a human designer can be learned, given that the system has access to sufficiently large and well-structured training data. The disadvantages, then, are that the learned information is not stored in a way that is intelligible to humans. It might be stored in the weights of a neural network, where no currently known approach is able to extract it from. This means that additionally, the decisions such a system reaches are opaque to both its users and designers.

The third family, then, comprises hybrid approaches. Hybrid approaches aim to overcome the pitfalls of top-down and bottom-up approaches by combining the strengths of both of them. They *represent* the normative information *symbolically*, while *learning* it using *machine learning tools*. This flexibility means that, so far, there is no clear trend emerging on either the symbolic representation to store normative information or the machine learning tools to acquire it. In Section 4.7, we will discuss hybrid approaches that are related to our framework in more detail. But first, we have to develop our

framework.

4.3. Training on Cases and Guessing Weights

The overall aim is creating a framework that can predict what the morally correct action(s) in a *case* are. As a reminder, a case represents a normative situation an agent could find herself in (see Definition 3.5.1). In it, she can choose between different actions, and each action might have reasons speaking for or against it which in turn might be modified or undercut.³ We represented this as a tuple (N, A) where N represents the normative situation as a Normative Graph and A is the set of permissible actions in this situation. That is, what we are after is a framework, that given a Normative Graph N , predicts the set of actions A .

In order to predict the set of morally correct actions, we need training material. This training material, our *training data*, takes the form of cases as well. We can think of these cases as situations where we know what the morally correct actions are. What remains to be done is to describe how we can go from the cases we train on to predictions in new cases.

Recall that we already talked about two ways cases allow us to determine the deontic statuses of actions in new cases. The first way was through the introduction of a consequence relation in Section 3.5.3. Here, we said that some set of cases $\mathcal{N}$ entails another case (N, A) iff for every possible weighting function w that satisfies every case in $\mathcal{N}$, w also satisfies (N, A) . The second way was when we showed that (in some cases) we can reverse engineer the weights of normative considerations, allowing us to trivially determine the morally correct actions in cases where we already have reverse engineered the weights of all considerations present in the cases.

In this chapter, we will combine these two insights and do something slightly more general (but less precise). Instead of looking at all *possible* weighting functions w , like we did for the consequence relation, we are trying to find a *single* weighting function w that best satisfies the cases in $\mathcal{N}$.⁴ We will attempt to find such a function, even if we cannot fully reverse engineer it through the set of cases. Our approach will thus be less precise than the reverse engineering we discussed in Section 3.5.2. This loss of precision, however, allows us, as we will show in Experiment 4.6, to deal even with inconsistent sets of cases as training material. Like earlier, the weighting function w we are trying to estimate depends on a function f determining the final weights of considerations and

³We also allow that modifiers and undercutters might themselves be modified or undercut and so forth.

⁴There might be multiple functions that satisfy a set of cases equally well. In such a case we settle for any of them. We define what makes a function satisfy a set of cases better than another, when we introduce the fitness score in Section 4.4.

a function g determining the deontic statuses of actions based on the final weight of reasons. To keep the function space tractable, we only consider a single function for g , namely the function g_1 (see Definition 3.4.13) where an action is permitted iff the sum of the reasons in favour of an action minus the sum of the reasons against this action is at least as large as that for all alternatives.⁵ How do we go about finding such a function w ? Here we turn to the *hybrid* part of our framework. We use genetic algorithms to infer w . This is what we will discuss next.

4.4. Genetic Algorithms

In this section, we will give a bird’s-eye overview of genetic algorithms, specifying the one we use to estimate weights.⁶ Genetic algorithms are a class of algorithms to solve optimization problems. In our case, what we are trying to optimize is the weighting function w . The more cases are satisfied by w , the better it is. For our purposes a weighting function is optimal, if it satisfies all cases (or the largest number of consistent cases, if the set of cases is inconsistent). Genetic algorithms were chosen because they are a good fit for this particular type of optimization problem. Remember that, as discussed in Section 3.5.2, a set of cases entails a set of (in)equalities, where a weighting function satisfies the set of cases if, and only if, it also satisfies every such (in)equality. Unfortunately, solving sets of inequalities without constraints is a hard problem and does not scale well with the numbers of inequalities in a system. So if we want to find an optimal weighting function for a set of cases, it is not efficient (or, often, feasible) to try to brute force a solution.

Genetic algorithms allow us to sacrifice precision for efficiency.⁷ Instead of attempting to find a solution to our problem directly, it makes use of (meta)heuristics inspired by natural selection to iteratively generate better potential solutions. The genetic algorithm does this via a population of *chromosomes* where each chromosome represents a potential solution to the problem. Each chromosome in turn is made up of a list of *genes* that represent the various variables of the problem. In our case, a chromosome is a weighting function w (defined on some number of considerations) and the genes of a chromosome are the individual weights of these considerations. Hence, each chromosome will have exactly as many genes as normative considerations we are interested in. This process of representing solutions as lists of genes is called “encoding”.⁸

⁵In preliminary tests, the choice of a function for g played a – perhaps surprisingly – small role.

⁶The source code of the implementation can be found at <https://github.com/DS-Thesis/PhD-Thesis-Chapter-4---Implementation/releases/tag/JurySubmission> on Github.

⁷We will see below in Section 4.6, that for quite a few cases, we will still be able to find an optimal solution.

⁸A quick note on the actual implementation: Instead of working with some list of considerations, we instead treat them (and all actions) as one huge graph. The cases we are interested in are then made up of *normative subgraphs*; that is Normative Graphs containing only some of the actions and

Any genetic algorithm consists of five steps. Step 1 is performed at the beginning, before the algorithm enters a loop. Steps 2–5 are then repeated until the optimal solution is found or the limit of steps is reached.⁹ Below, we present the general idea behind each step and also comment on the way each of them works in our implementation.

Step 1: Initialization The population is initialized with an initial value for its genes. In our implementation, this value belongs to $[0, 5]$ and is initialized to the median value, namely 2.5.¹⁰

Step 2: Evaluation A fitness score is computed for every chromosome. Here, the fitness score is simply the number of cases in the training data that (the weighting function represented by) the chromosome satisfies.

Step 3: Selection A new population of chromosomes (children) is generated from the previous ones (parents). There are several selection methods, but the main idea behind them is the same: the probability of a chromosome being preserved for the next population is proportional to its fitness score. We use the *steady state* selection method, which differs from other methods in that it does not necessarily replace all members of the population with their children. Instead, it replaces the parents only in case the children are performing better than the parents did (unless the newly generated children duplicate existing chromosomes).

Step 4: Reproduction (or Crossover) Children are generated from the parent population, replacing the parents. Usually, two parents are selected, and two of their children are produced, each having a mix of the genes from both parents. This may affect only a part of the total population. We use the *single-point* crossover method. This method generates two children. The parent chromosomes are split in two, with each child inheriting half a chromosome from each parent (see Figure 4.1). This is a simple reproduction method that is computationally inexpensive.

Step 5: Mutation For each chromosome, a different value replaces some genes. Our mutation operator consists in randomly selecting one weight of a consideration and changing its weight to a random floating-point value in $[0, 5]$. This operator has the advantage of being time-efficient, and prior testing has shown that the desired result can be reached without relying on more sophisticated functions. Choosing a random value

considerations of this big graph. Furthermore, to make the implementation and computing easier, we take the weights of the considerations as the weight of edges in the (directed graph of the) Normative Graph instead of the nodes themselves.

⁹If there is no optimal solution – say, because of noisy data – it is possible to stop the algorithm and get its current best solution as the final output.

¹⁰The weights we give to these considerations are, for implementation reasons, rational numbers in the bounded interval $[0, 5]$ with a precision of 0.1. We do this to keep the computational cost low and because we think that the interval $[0, 5]$ is a good fit of the difference in strength one can find *in the wild*. In any case, preliminary tests suggest that both the precision and the exact interval used has little impact on the predictions of our algorithm.

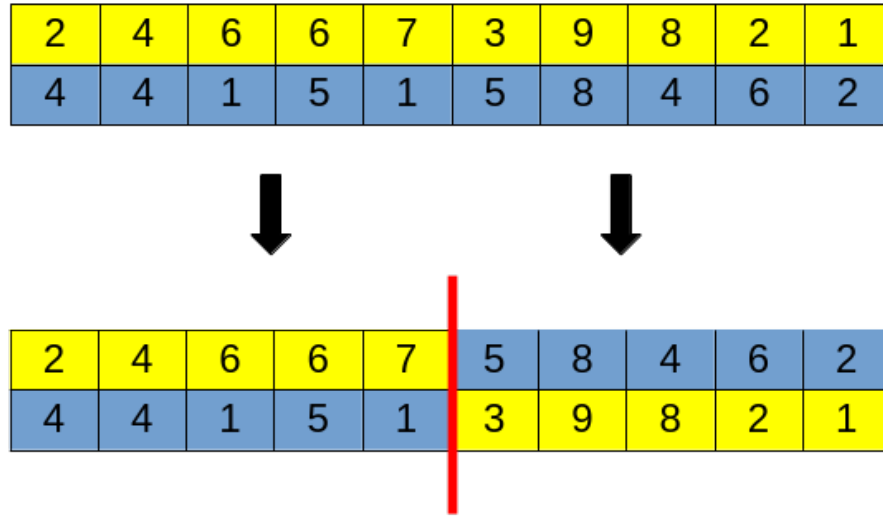


Figure 4.1.: Example of a single point crossover.

instead of moving to neighbour values has a clear advantage: it helps to get out of local optima more easily.

This algorithm gives us a computationally inexpensive way to estimate a weighting function w based on a set of cases. As we will see below in Experiment 4.6, the prediction accuracy we get this way is quite high, and we have no need for more sophisticated (and computationally expensive) algorithms.

With the framework now in place, we need to evaluate it. This is the goal of the next two sections. First, we discuss how this is done, and second, we present some experiments.

4.5. Evaluating the Framework

4.5.1. Ideal Scenario

Imagine for a moment that we had access to a moral oracle that could reliably tell us what the morally correct actions in some situation are.¹¹ Then the way to evaluate our

¹¹We phrase this section in the language of moral realism. We do this because we think reasons-talk lends itself easily to realism, but, strictly speaking, everything we present in this section can be

framework would be simple, as long as we are correct in our underlying assumption that every moral case can be modeled as a Normative Graph. We could give the moral oracle a set of cases and ask it for the morally correct options in each case. We could then train our framework on some fraction of these cases (say 70%) and evaluate for what percentage of the remaining cases we can correctly predict the morally correct actions. If the percentage is high, our framework provides accurate predictions, if it is low, our framework does not provide accurate predictions.¹²

Unfortunately, moral oracles, at the time of the writing of this thesis, are not readily available, and so we need a new approach. We will imitate what a moral oracle does using *synthetic data*.

4.5.2. Synthetic Data

To illustrate how we generate this data, let's assume, for now, that any moral scenario can indeed be represented in our model. That is, every scenario can be represented as a Normative Graph, that is, as a set of reasons, modifiers and undercutters which come equipped with a default weight and that this default weight in turn generates a final weight via one of the functions $f_1 - f_4$ and the final weights of considerations ultimately determine the deontic status of actions via the function g_1 .

As the algorithm we employ trains on cases, the only unknown element is the concrete weight of the normative considerations found in those cases. That means that the only ground truth a moral oracle would need is the weights of these considerations, it can compute the rest. And this is in effect what we will be doing.

Our synthetic data is generated from a set of considerations that we supply some random weights to. We can think of this set of considerations as representing some well-circumscribed area of moral discourse. Every consideration having a potential impact on decision-making will be included in this set. Next, we generate a bevy of Normative Graphs containing some (smaller) subset of the set of considerations along with a set of possible actions. For all these graphs, we then first determine the final weights in each graph (given a function $f \in \{f_1, \dots, f_4\}$) and then the deontic statuses of the action in the graph (given the function g_1).

These Normative Graphs and the deontic statuses we determined then serve as our

rephrased to fit with a non-realist position as well.

¹²There is an alternative way to test our framework, if the moral oracle were to give us the correct *weights* of reasons (and other considerations) as well. We could train it on a set of cases and then evaluate how closely the weighting function w we estimated matches the true weights of reasons and other considerations. Comparing different weighting functions, however, is not a trivial task, as we have seen that for g_1 (and many of the functions $g_2 - g_6$), the weight of reasons is a purely relative measure and finding a similarity metric for these two functions is quite challenging.

4. Machine Ethics from First Principles

cases. In most of the experiments we discuss below, we simply split these cases into our training data on which we train the genetic algorithm and our validation set on which we evaluate how well its predictions match the cases in this set. For some experiments – we will explain this in more detail below – we will create the training data and validation set independently.

4.6. Experiments and Empirical Results

In this section, our goal is to evaluate the algorithm empirically. To this end, we perform four experiments.

To test the algorithm, we need two things: training data and a validation set. Here we use *synthetic data* for both, which we derive as described above. The number of considerations was manually defined for each experiment, and the Normative Graphs we generate have a maximal graph depth of 49 considerations in every experiment. The experiments were run with a Python implementation, using the PyGAD library, and on a machine equipped with an 11th Gen Intel Core i9-11950H at 4.80GHz and 32Gb of RAM. While our implementation uses a single thread, our approach can easily be multithreaded to iterate faster over the training data. The population contains 10 chromosomes, and we perform a crossover on 40% of the population. All the generated graphs for our cases contain 5 actions, and we perform 100 steps (i.e., generations) for the training phase. The number of considerations in each graph depends on the test. In each run, we calculate the set of permitted action(s) for 500 cases with 40% (200 cases) serving as the training data and 60% (300 cases) serving as the validation set.¹³

Prediction Accuracy

The first experiment tests for the effects of the number of normative considerations we consider on the algorithm’s accuracy on the training data and the validation set, and the time and steps the algorithm needs to find a solution.

To test this, we used the Additive Function (see Definition 3.4.6) to create both the training data and validation set. We run the algorithm on sets of consideration varying in size between 5 to 100. The results for each size were averaged over 20 runs.

Our starting hypotheses for this experiment were the following: since we use the same size of training data (200 cases), the accuracy will begin to drop with larger sizes as

¹³We vary this approach slightly for the second experiment, which we will discuss below.

the training data covers less and less of all possible cases.¹⁴ For the time it takes the algorithm to run, we expected a polynomial increase with the number of considerations, as it is based on the grounded extension algorithm proposed by (Nofal et al., 2021).

In Figure 4.2, we plot the accuracy of our prediction on the training data and the validation set over the number of considerations. In Figure 4.3, we plot the time it takes for one run to finish and the number of steps it takes for the genetic algorithm to stabilize on a solution.

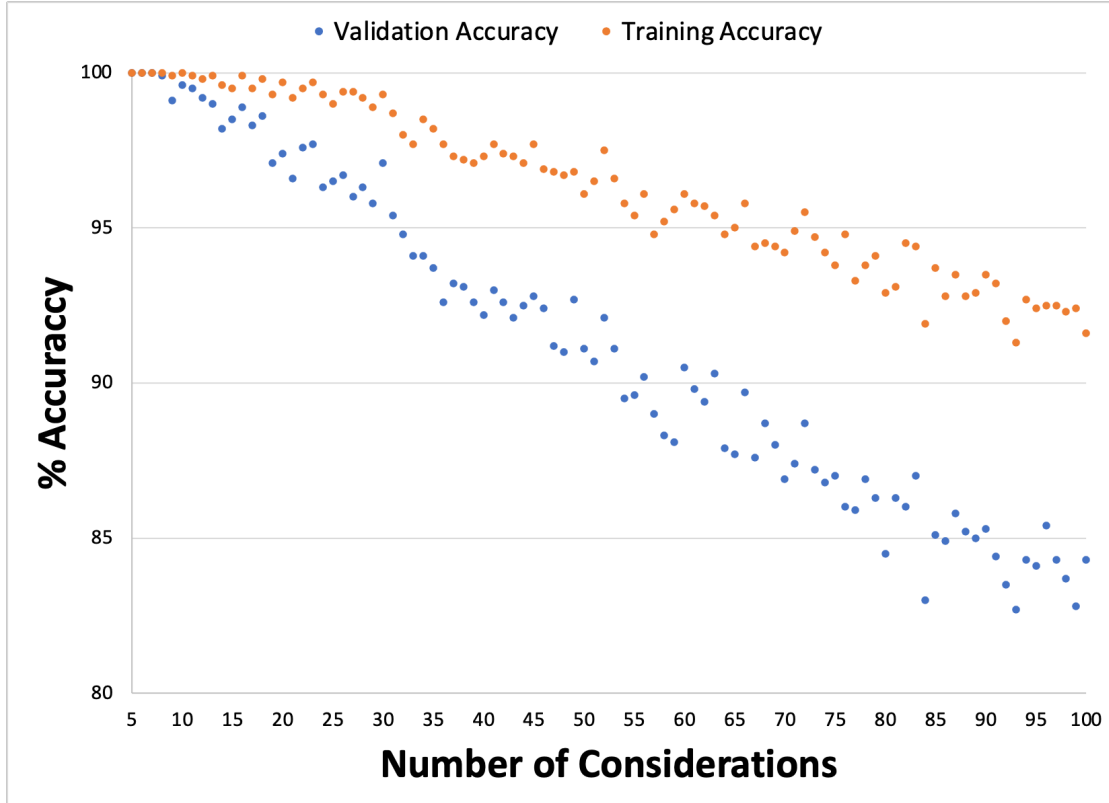


Figure 4.2.: Accuracy

The data supports our accuracy hypothesis. The accuracy of predictions begins to gradually drop, but stays roughly at 90% accuracy (on the validation set) for 50 considerations. Even for 100 considerations, when the algorithm is trained on only a small fraction of possible cases, the accuracy stays well above 80%. The time it takes for the algorithm to run remains roughly linear for the graph sizes we use, while the number of steps it takes to reach a solution appears logarithmic.

¹⁴How many cases can be created from n considerations is highly dependent on how they are structurally related. On the one extreme, if they form a graph with no modifiers or undercutters, there are 2^n possible cases, on the other extreme, if they form a graph that consists only of a single chain of one reason and modifiers and undercutters, there are only $n + 1$ possible cases.

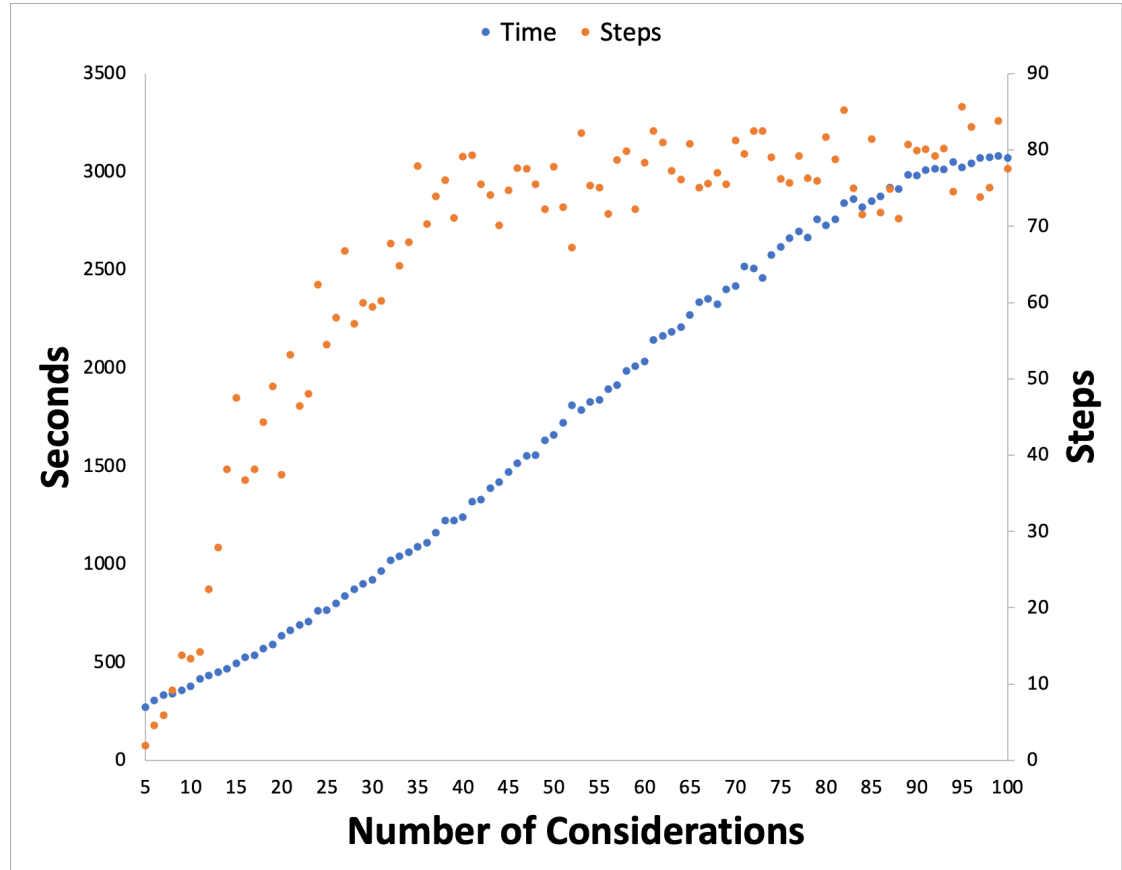


Figure 4.3.: Time and Steps

Importance of the Choice of Function

The second experiment tests whether there is a difference in accuracy between the four functions f_1 – f_4 .

To test this, we analyzed the accuracy of each function f_1 – f_4 when trained on four sets of synthetic data. One set was generated by the function we were testing, while each of the other three was created by one of the other three functions. Then we compared the accuracy of the function on the validation data from each set. The graphs contained 50 considerations, and the results were averaged over 20 runs.

We had two hypotheses for this experiment: First, that each function would show better accuracy on its own dataset than on any other set, and second, that some functions will perform better on data not created by that same function.

	f_1	f_2	f_3	f_4
$\mathbf{f}_1$	90.3 ± 3.7	94.5 ± 2.7	95.4 ± 2.6	94.9 ± 2.8
$\mathbf{f}_2$	93.3 ± 3.3	92.1 ± 4.7	94.9 ± 2.5	95.7 ± 2.3
$\mathbf{f}_3$	92.7 ± 3.3	94.7 ± 2.4	94.7 ± 2.6	95.2 ± 2.6
$\mathbf{f}_4$	93.0 ± 3.0	94.4 ± 3.3	95.2 ± 2.5	94.6 ± 2.4

Table 4.1.: Each row presents the accuracy on the validation set in percent when trained (and validated) on data created using the function given in the column. We give the standard deviation σ as $\pm\sigma$.

The results, displayed in Table 4.1, partly confirm and partly refute our hypotheses. The Additive Function, f_1 , performs *better* when used on data not created by itself, while the other functions perform at least as well. For the mixed functions f_3 and f_4 (and to a slightly lesser degree for the multiplicative function f_2), it appears not to matter much which function is used to create the synthetic data. These results suggest that, for designing a tool for moral learning, the (rather simplistic) Additive Function might be an ill fit, but the other functions perform better and are roughly on par with each other. We take this to be good news.

Noise Robustness

This experiment tests the effects of noise in the data on the accuracy of the algorithm. It is important insofar as real-world applications might require working with imperfect data. For instance, such data might contain cases where forbidden actions are taken to be permissible, or it might even contain inconsistencies. Hence it is important for any machine ethics approach to deal with this kind of noisy data gracefully. To test this, we gradually introduce noise to the training data and assess the accuracy.

We created noisy data as follows: for each case in the training data, with a certain probability, we replace the actual set of permissible actions with a random set of actions. We tested this for 10%/20%/30%/40% probability on a set containing 50 considerations. The results were averaged over 20 runs.

Our hypothesis for this experiment was this: with a small amount of noise, the algorithm will remain accurate, performing better on the validation set than on the training data, but, with a large amount of noise, we see a rapid drop in accuracy.

The results, summarized in Table 4.2, partly confirm the hypothesis and partly refute it. When the training data contains some noise, the accuracy remains high at between roughly 90% for the additive and almost 95% for the mixed functions. However, even when 40% of the training data is noise, the accuracy never drops below an average of

4. Machine Ethics from First Principles

	10%	20%	30%	40%
$\mathbf{f}_1$	91.1 ± 5.3	90.5 ± 4.2	90.4 ± 3.2	90.6 ± 3.5
$\mathbf{f}_2$	93.4 ± 3.8	92.7 ± 3.3	93.1 ± 3.0	92.6 ± 3.8
$\mathbf{f}_3$	94.6 ± 3.7	94.0 ± 3.1	95.3 ± 2.3	94.7 ± 2.8
$\mathbf{f}_4$	94.3 ± 3.8	93.0 ± 4.6	94.3 ± 3.1	94.2 ± 2.6

Table 4.2.: Accuracy for the different evaluation functions on the validation set given noisy training data. We give the standard deviation σ as $\pm\sigma$.

90% and even stays at roughly 95% for the two mixed functions.

We interpret these results as follows: first, the fact that accuracy remains high with noise in the data suggests that the algorithm is fit to be used on real-world data. However, it might seem mysterious at first glance, how high the accuracy is even when 40% of the training data is random noise. We conjecture that the accuracy does not come from the estimates of the weights of considerations alone, but also – perhaps primarily – from the fact that the cases are encoded in a graph structure. Hence, in the next experiment we try to quantify the effect that structure has on accuracy, independently of any weight estimation using the genetic algorithm.

Effects of Structure on Accuracy

To test the effect of the normative structure alone, we created the validation set as usual, but, instead of running the algorithm on training data, we simply stipulated that it assigns the weight of 2.5 to all considerations. Then we used the validation set to check the predictive accuracy of these constant weights.¹⁵ We ran this test using all four functions and averaged over 1000 runs.

Since we observed high accuracy even with highly noisy data, our hypothesis for this experiment was that the structure is responsible for at least 50% of the accuracy.

f_1	f_2	f_3	f_4
62.0 ± 7.8	69.1 ± 8.7	72.6 ± 8.3	72.5 ± 8.2

Table 4.3.: Accuracy in percent when using constant weights. We give the standard deviation σ as $\pm\sigma$.

The results, shown in Table 4.3, fully confirm the hypothesis. Using constant weights results in about 70% accuracy for every function, except for the additive one, which, like in previous experiments, performs significantly worse.

¹⁵We chose 2.5 since it lies halfway between the smallest and greatest value we allow for weights. Preliminary tests have shown that other values (except for 0) produce similar results.

We interpret these results as follows: The information contained in the structure of a Normative Graph is the main carrier of normative information. In fact, in many cases the structure is enough to determine the morally correct actions with some confidence. This in turn suggests that the default weights of considerations – much like the choice of the functions for calculating final weights – are less important than one might have initially thought. Getting good default weight estimates accounts for only roughly 30–40 percentage points of the accuracy of predictions.

4.7. Related Work

In this section, we compare our framework to related work. We compare our work to what we take to be the two most similar approaches to ours, (Freedman et al., 2020) and (Anderson and Anderson, 2018) as well as to (Canavotto and Horty, 2022), who, like us, propose a framework based on reasons. We will highlight specifically what we take to be the advantages of our approach.

4.7.1. Freedman et al.

Freedman et al. propose a hybrid framework whose goal is to automatically distribute human kidneys to potential transplant recipients in so-called kidney exchanges in a morally acceptable way. Kidney exchanges work, roughly, as follows. A patient in need of a kidney has a potential donor, usually a family member or friend, who is willing to donate their kidney. Very often, though, patient and potential donor are not histocompatible. Since, unfortunately, the pool of people in need of kidney transplants is high, we find many of these incompatible pairs. Sometimes, for two pairs, the patient of the first pair will be histocompatible with the donor of the second pair and vice versa. Then, the two patients can exchange donors. There can also be more complex relationships of histocompatibility, such as when the donor of pair 1 is compatible with the patient of pair 2 whose donor is in turn compatible with the patient of pair 3 whose donor is in turn compatible with the patient of pair 1, forming a cycle of compatibility.¹⁶

The best way to match these pairs is both a computationally expensive (see (Biro et al., 2009)) and morally laden question. It is in this context that Freedman et al. propose their machine ethics framework. To give an example where the donor-matching might not be quite controversial, consider the example found in (Freedman et al., 2020). You have one patient-donor pair that can be matched with two potential other pairs. The patient in one of these other pairs is old, has a bad health outcome prognosis and is a

¹⁶The logistics of ever-increasing cycles become more difficult to manage, though. Impartial, altruistic donors (those who do not donate a kidney merely to save some specific person close to them) also play a big role in these exchanges, but we skip over this complication here.

4. Machine Ethics from First Principles

heavy drinker, while the patient in the other pair is young, has a good health outcome prognosis and rarely drinks.¹⁷ Then, all other things equal, we should match with the second pair rather than the first.¹⁸

Freedman et al. now propose a way to match the patient-donor pairs (and thereby allocating kidneys) in a morally correct way. They proceed in two steps, first they want to maximize the number of people receiving kidneys (and they do this with standard graph theoretic tools), but there are potentially multiple ways to achieve the same number of transplants.¹⁹

The way they choose which pairs should get matched and hence which patients should get the kidneys will sound quite similar to our model. First, as they only consider three independent two valued variables (prognosis good/bad, drinking often/rarely, young/old), this gives them a total of $2^3 = 8$ *profiles* that can be compared. They then assign weights to these profiles that, roughly, correspond to whether (and how much) a certain profile is preferred over another when considering whom to assign a kidney. However, as they later discuss in the paper, the concrete way to assign weights does not seem to matter all that much, and we omit the details here.²⁰

We think that assigning weights to these profiles corresponds, very roughly, to what we do when we compare actions in a Normative Graph having already summed up all the weights of the reasons speaking in favour of and against each action. Freedman et al. also discuss ways to generate weights for the individual factors (e.g. old age, drinking heavily, rarely drinking, etc.), which, we think, corresponds rather nicely to weights of reasons in our model. Notably though, the way they determine the weights of the individual features is always from the values assigned to each profile, and not, as in our framework, directly from the training data.²¹ Again, the precise way to determine these weights does not seem to matter a lot for the prediction accuracy of their framework. Lastly, in order to determine which pairs get matched (and thus which patients get kidneys) Freedman et al. simply use the matching where the sum over all the weights of patient profiles involved is maximized.

¹⁷We replaced the factor “skin cancer in remission” as it appears in Freedman et al.’s paper with “bad health outcome” because we think it makes the point clearer.

¹⁸This, at least, seems to be the consensus of the clickworkers on Amazon’s Mechanical Turk that were surveyed for the paper.

¹⁹Nothing in the framework they present hinges on the fact of first maximizing the number of transplants. They chose this “[b]ecause [their] human subject data and analysis do not involve comparisons between different quantities of patient profiles” (Freedman et al., 2020, p. 7).

²⁰(Freedman et al., 2020) because they use survey data for the question of who should get a kidney, use Bradley-Terry scores (see (Bradley and Terry, 1952)) to determine these weights. However, they also consider other weighting schemes, and it doesn’t seem to matter all that much which scoring system is used.

²¹Specifically, they propose to generate these values such that the value of each *profile* is a linear combination of its constitutive factors. That is, it is a weighted sum (plus one error term that results from their use of statistical methods to determine these values).

Next, we compare their framework to ours. There are some obvious similarities between the two, but also considerable differences. The first obvious similarity is that in both frameworks, the morally correct actions are determined by numerical weights on normatively relevant features. In our case, these are the (final weights of) reasons, in Freedman et al.'s case they are either complete profiles or the individual features they are composed of, like old age or heavy drinking. In Freedman et al.'s case, these weights are generated using statistical tools, whereas we use a genetic algorithm to attempt to find an optimal solution. While there are profound differences in these tools, we think the overall idea is still quite similar. Collect training data, extract the weights, and then use those weights to determine what ought to be done in future cases.

Freedman et al. collect the training data by asking clickworkers to determine who would get a kidney in a hypothetical scenario. In general, they seem to think that a kind of public participation (perhaps guided by experts) is a good way to collect the training data. We do have some reservations about treating moral decision-making as an essentially democratic enterprise, but this only concerns the way training data is collected, and this is a question entirely independent of the workings of the framework once this data is in place – both for Freedman et al. and us.

Still, we think there are some shortcomings in Freedman et al.'s framework that we do not share. First, their model is essentially flat, allowing for no distinction between modifiers, undercutters and reasons. This is not necessarily a disadvantage, as long as the model can still represent the underlying normative structure in some way and hence provide accurate and reliable predictions. Whether Freedman et al.'s model satisfies either of these is dubious. At least as it stands now, the paper does not provide any way to test the accuracy or validity of the framework against an external standard. We also do not believe that the choice of representing the choiceworthiness of profiles as a derived statistical artifact is immediately (meta)ethically plausible. In all three cases, we believe our model tells a more convincing story.

Our model also has a further advantage. In general, we think our framework can deal with sparse data better than Freedman et al. In effect, in order to be able to decide between two different actions (that is, patient profiles), they need to have that exact patient profile in their training set at least once (and, arguably, for the system to work in practice, they would need to have a comparison between *each combination* of profiles in the training data). We, on the other hand, only need each *factor* (that is, reason, modifier or undercutter) to be present. This is a big difference between having profiles (which, translated to our model, would roughly correspond to the entire Normative Graph). We thus think that our framework allows us to work with less training data and can thus be used in cases where it is not feasible to collect training data on every possible combination of factors.²² If for example, we used only 10 attributes that have three

²²One way for Freedman et al. to address this is to use fewer cases and use those to give weights to the individual factors they are composed of. This is a possible response, but we do not think it is a

4. Machine Ethics from First Principles

values, low, medium, and high (for example: young, middle-aged, old, or non-drinker, medium drinker, heavy drinker), this would lead to $3^{10} = 59,049$ possible profiles, more than can be easily surveyed or encoded by hand. If, additionally, we would want to have a comparison between *each* of these profiles, we would need to include at least $\binom{3^{10}}{2} = 1,743,362,676$ (that is, almost 2 billion) cases in the training data. In our framework, this only amounts to 30 normative considerations that need to be included in at least one case each. We have seen that for 30 considerations, we can still achieve a prediction accuracy of roughly 95% with just 200 cases as training data even for the Additive Function which has the worst prediction accuracy of all the ones proposed. Notably, this performance improvement is not solely due to the graph structure we impose on our cases but also because our approach allows us to directly give weight to the individual factors (i.e. considerations) instead of having to take the detour of first comparing profiles.

Overall, we think these advantages mean that our model is both more general, stands on more secure philosophical footing, and is more computationally efficient.

4.7.2. GenEth

Next, we compare our framework to the *GenEth* framework due to (Anderson and Anderson, 2018). GenEth relies on (manually selected) “ethically relevant features” which play a similar role to our normative reasons. Like our framework, GenEth is trained on a set of cases for which the morally correct action(s) are known. A case is completely characterized in the GenEth framework by how much it violates or promotes these ethical features, and can be represented as a tuple of integers (usually taken from $[-2, 2]$). Each position in the tuple corresponds to an ethically relevant feature, and the integer indicates the degree to which the action violates or promotes this feature. To take a simple example from medical ethics; assume a doctor needs to decide between two actions a and b while treating her patient and only two ethically relevant features have an import for her decision, harm coming to her patient and her patient’s autonomy being respected. Action a is, say, giving the patient a placebo so that they do not self-medicate in a harmful way. This action, let’s assume, fully prevents harm to the patient, but does not fully respect their autonomy. So we can represent this action as the tuple $(2, 1)$. Alternatively, for the action b , the doctor could simply tell the patient the truth and not give them any medication, knowing full well that once the patient arrives at home they will immediately take copious amounts of painkillers, potentially damaging their liver and kidneys. This action can be represented as $(-1, 2)$. The GenEth framework can include as many ethically relevant features as desired in the domain of application, as long as each be represented by their degree of violation/promotion for any action one might take.

particularly good one, as the way these factors are given weights depends on the weights of profiles, we believe to give an accurate number to them, one would still need a large training dataset.

With this representation in place, the GenEth framework then trains on a set of cases. Cases in the GenEth framework are choices between two actions (i.e. tuples of integers) where one of the actions is favoured over another. (Anderson and Anderson, 2018) suggest that domain experts can answer which action is preferable for some of these pairwise choices.

Internally, these pairwise choices are represented in a propositional language (with $\leq$ and $=$ symbols).²³ If, for example, we should prevent harm to the patient over (fully) respecting their autonomy (that is, we should choose action a , represented by the tuple $(2, 1)$ over action b , represented by the tuple $(-1, 2)$, GenEth first calculates the differences between the ethically relevant features, harm, d_1 , and respecting the patients' autonomy, d_2 , of the two actions. That is, we are calculating $\Delta d_1 = 3$ for harm and $\Delta d_2 = -1$ for the patient's autonomy. In this case, we know that a is preferred over b , which we will write as $p(a, b)$. We can express this case as a single propositional formula $(\Delta d_1 = 3 \wedge \Delta d_2 = -1) \rightarrow p(a, b)$.²⁴

This representation of cases allows GenEth to make use of inductive logic programming, which, roughly, extracts the most general principle sanctioned by the cases in the training set. For example, if we added some more cases, it might generate a principle that says that if an action that strongly violates a patient's autonomy and only moderately prevents them from being harmed, then, regardless of what other ethically relevant features are in play, an action that respects a patient's autonomy is always preferred as long as it does not harm the patient severely. This principle can then be applied to new cases to determine the morally correct actions in these cases.

Like with Freedman et al.'s approach before, there are some obvious similarities between our approach and GenEth. We will first briefly discuss the similarities – this will closely mirror our discussion of Freedman et al. – and then discuss the differences in more detail, highlighting what we take to be the advantages of our approach.

First, the “ethically relevant features” GenEth uses, we think, closely resemble the role reasons play in our model.²⁵ Like in our framework, these are gradable (albeit to a much lesser extent). And lastly, a set of cases where the morally correct action(s) are known is used as training data, which then allows GenEth to predict the morally correct action(s) in new cases.

²³While it might seem that we are doing arithmetic here, note that we are not. We have a limited number of propositional constants, a small set of integers, and a relation on these integers. This can easily be translated back to a propositional base logic.

²⁴Note again, that even though the formula does not look like propositional logic, as we are dealing with a finite set of ethically relevant features and differences, we can easily translate them back to propositional logic.

²⁵Anderson and Anderson use Rossian Pluralism as a guide. That is, the promotion of ethically relevant principles is seen as a *prima facie* duty. We don't think we do the Rossian position harm if we take it to be easily translatable into talk of reasons.

4. Machine Ethics from First Principles

Like Freedman et al. the representational structure of GenEth is *flat*: there are no differences between types of normatively relevant considerations.²⁶ We believe that this is a disadvantage of their model, but it isn't as limiting as in the case of (Freedman et al., 2020), as GenEth at least has the connectives of propositional logic to allow for some structure between the features.

Additionally, by necessity, GenEth has to limit how fine-grained ethically relevant considerations can be represented. Every action must be able to be classed into categories of how much it violates or promotes some ethically relevant features (usually represented as integers in $[-2, 2]$). This is because otherwise, the amount of training data needed to be able to extract principles would soon become quite large. Our model, on the other hand, can be as fine-grained as we need it to, the only constraint is the computational resources we have available.

There is another problem with GenEth that it is worth briefly discussing. GenEth implicitly assumes that this representation is adequate, even omitting the problem of neatly categorizing how much an action violates or promotes some ethically relevant feature. That is the problem that the principles in GenEth are expressed using the integer difference in ethically relevant features. So for example, if we have the ethically relevant feature of promoting well-being, and we would like to express the principle that any action that promotes well-being more than another (without regard for any other factors) is preferred, we could express this principle in Anderson and Anderson's style as $\Delta d_w \geq 1 \rightarrow p(a_1, a_2)$ where Δd_w is the difference between the well-being between action a_1 and a_2 . So for example, if action a_2 does not promote well-being at all (it gets assigned the integer -2 for example) and action a_1 promotes it just slightly better (it gets assigned the integer -1), then $\Delta d_w = 1$. But equally, if action a_1 promotes well-being substantially (say, $+2$) while a_2 only somewhat (say, $+1$), $\Delta d_w = 1$. That is, there is no way for the principles on which GenEth relies to distinguish these cases. But we have seen that one could plausibly be a *welfare satisficer*, saying that as long as an action promotes well-being somewhat, it is okay to perform it, but if it does not, then one should always prefer the action that serves well-being the most. That is, we would like to distinguish the first from the second case. Unfortunately, at least as GenEth is introduced in (Anderson and Anderson, 2018), this distinction is simply not possible. Our framework does not suffer from this flaw.

In the case of ethical learning, GenEth's use of inductive logic programming allows it to output ethical principles encoded in a logical language. This is something our model – without supplementing it with a logical language – cannot do. These principles might be somewhat complicated, but, if there are not too many ethically relevant features in play, they are easily human-interpretable.

²⁶Although (Anderson and Anderson, 2018) mention “meta-features” in passing – their examples are time and probability – they also suggest that these shouldn't be part of the representation of the normative structure (see pp. 342–343).

Still, relying on inductive logic programming has a disadvantage as well. When applied to a set of cases, its goal is to find a general principle satisfying all cases in the set. This has the problem that if the set is noisy or even inconsistent, this method invariably fails. This challenge might be overcome with further tools, but as it stands now, it cannot deal with noise and inconsistencies as gracefully as our framework can. As we have seen in Experiment 4.6, our framework is able to generate reasonable predictions even with very noisy (or inconsistent) data.

4.7.3. Canavotto & Horty

The last framework we look at, (Canavotto and Horty, 2022), is not as closely related to ours as the two previous ones when it comes to the way normative information is learned, but its underlying model, like ours, is based on reasons, and it very explicitly learns on sets of cases whose representation is similar to ours.

Canavotto and Horty’s model is inspired by case-based reasoning in common law where, roughly, judges have to decide cases according to the *precedent* set by previous rulings. The basic set-up of the model is this. In a case, a judge has to decide either in favour of the defendant (represented by the letter δ) or in favour of the plaintiff (represented by the letter π). She has to do this on the basis of the (legally relevant) factors of the case, that Canavotto and Horty assume can be partitioned into factors favouring the defendant F^δ and factors favouring the plaintiff F^π . Let us call elements of the set F^δ $f_1^\delta, f_2^\delta, \dots$ and elements of F^π $f_1^\pi, f_2^\pi, \dots$. When a judge is confronted with a case, what she does is decide based on a subset of the factors in play and decides that a specific subset of these factors decisively favours either the defendant or plaintiff. Canavotto and Horty call this the *rule* of a case, represented as $\{f_1^\delta, \dots, f_n^\delta\} \rightarrow \delta$ if it is in favour of the defendant or $\{f_1^\pi, \dots, f_n^\pi\} \rightarrow \pi$ if it is in favour of the plaintiff. The case can then be represented as the triple (X, r, s) , where $X = F^\delta \cup F^\pi$ is the total set of factors in a case, r is a rule as described above and s is the one in whose favour the rule decides. r has to be such that the set of factors on which the rule is based is included in X and s has to be ruled in favour of by r .

Each such case then returns a priority ordering on *reasons* in Canavotto and Horty’s terminology. Here reasons are understood as sets of factors for either the defendant or the plaintiff. The best way to think about this is again as a legal model. If the judge rules in favour of the defendant based on some set of factors, then these factors are better than every possible combination of factors the plaintiff could have (or did) bring. With the same line of thought, if the defendant brings even more factors in their favour to the table, this set of factors will be at least as strong. This gives us the following definition:²⁷

²⁷We slightly adapt the notation used in (Canavotto and Horty, 2022).

4. Machine Ethics from First Principles

Definition 4.7.1 (Priority Ordering Derived from a Case). *Let $C = (X, r, s)$ be a case, and let U and V be reasons such that the underlying factors of V favour s and the underlying factors of U favour their opponent. Then the relation $<_C$ representing the priority ordering on reasons based on the case C is defined by stipulating that $U <_C V$ iff $U \subseteq X$ and the factors present in the rule are a (non-strict) subset of V .*

This priority ordering can be lifted to the level of sets of cases easily. Let $\mathcal{M}$ be a set of cases. Then $U <_{\mathcal{M}} V$, iff there is a $C \in \mathcal{M}$ such that $U <_C V$. This definition is then used to define a *consistent* case base as a set of cases $\mathcal{M}$ where for no two reasons U, V it is both the case that both $U <_{\mathcal{M}} V$ and $V <_{\mathcal{M}} U$.

With these preliminaries out of the way, we can state the central idea in the framework. A judge is allowed to rule in favour of the plaintiff or the defendant only as long as that is consistent with precedent. This means that, given a consistent case base $\mathcal{M}$ and some legal factors X , the court can only choose a rule r such that $\mathcal{M} \cup \{(X, r, s)\}$ remains consistent.

Canavotto and Horty suggest that this framework can be applied outside the legal domain and propose one area that we have already discussed and that is worth analysing in more detail: kidney exchanges. The main idea is treating two patients competing for a kidney as analogous to a plaintiff and a defendant. The facts about their respective cases are treated as factors in favour of either them or the other patient. The fact that a patient is a heavy drinker, for example, is a factor in favour of the other patient. A rule is then a decision about who should get a kidney in a choice between two patients based on the particulars of their cases. Once a number of these cases have been added to a case base, in some cases, who should get a kidney then follows from these cases, in others there is no precedent.

This framework has some striking similarities to our approach, but also some obvious differences, on which we will focus next. The first such difference is that Canavotto and Horty do not attempt to first construct an underlying moral theory before applying it to new cases. They operate purely on the level of cases. The only substantive notion of morality is one of consistency. Contrast this with our model or (Anderson and Anderson, 2018). In our model, we aim to identify every morally relevant feature (that is: all relevant considerations and their weight) first and then apply them to new cases. In the case of (Anderson and Anderson, 2018), they aim to first find the most general moral principles in play and then apply these principles to new cases. For now, we want to remain agnostic whether this is the right approach or not, but instead focus on whether Canavotto and Horty's approach is really all that different from what we are doing.

Imagine for a moment that the training material for our framework is consistent (in the sense defined in Definition 3.5.3). Our approach is then to find some weighting function w that satisfies all cases and then use this weighting function to predict the

morally correct actions in further cases. This means that we do not simply allow any case consistent with our set of cases, but only those that are (additionally) satisfied by the weighting function w . Since there may be more than one weighting function w that satisfies the set of cases, what actions we deem morally correct in a case depends on the choice of w . However, if some case is *entailed* by a set of cases, the outcome will be the same for any choice of w .

This means that in a way, our model is less liberal than that of Canavotto and Horty. In our model, we will claim that some set of actions is (im)permissible purely based on the weighting function that was chosen. Had we chosen another weighting function, we might have returned a different set of actions. Note though, that whatever set of actions we claim are (im)permissible, we will never make inconsistent claims as, by definition, there is a single weighting function w that governs which actions we deem morally correct. This means that, by definition, any arbitrary set of cases will always be satisfied by the same weighting function and thus consistent.

There are other differences that we have talked about when discussing the other two frameworks and will only briefly mention here. First, Canavotto and Horty’s model is *flat* and does not allow for factors to play the role of attenuators or undercutters. We believe this is a considerable downside when the model is applied broadly as there is no way to account for the interaction between normatively relevant considerations without treating every amplified or attenuated reason as a primitive factor in the model. This, we believe, would not scale very well. Second, Canavotto and Horty’s model does not make use of any machine learning or other (quasi-)statistical tools. Instead, it constructs the priority relation between reasons in a rather straightforward manner. We have not analyzed the complexity and computational difficulty of computing this relation in detail, but we believe that our model, as mentioned earlier, scales better with the number of normative considerations involved. Canavotto and Horty have to calculate a priority ordering for every subset of factors (that is, the reasons) which means that for n factors, they need to compare at worst 2^n different reasons.²⁸ Further tests are needed, but we feel confident that the results of Experiment 4.6 provide evidence that our model will prove more efficient.

4.8. Conclusion

In this section, we briefly outline possible future research and conclude. The first and most obvious future work lies in testing the framework we have proposed in real-world contexts. There are two challenges here. The first challenge lies in representing norma-

²⁸To be clear, this number can be lowered with efficient algorithms as, if you know that some set of factors U is ranked higher than some set of factors V , every superset of U will also be ranked higher than V .

4. Machine Ethics from First Principles

tive scenarios as graphs, and the second challenge is to collect the training data. Due to the information stored in the structure in our framework, the first challenge is likely the more important one. We need to transfer real-world scenarios (or, better, descriptions of such scenarios) into Normative Graphs in a reproducible and adequate way. We believe that the toolkit of the social sciences can offer guidance here. After a training session, we can use coders to identify the structures of real-world normative situations. As it might be controversial what these are (even between philosophers), we believe the best way forward is to have a (possibly opinionated) philosopher provide examples of what normative features are relevant. Next, of course, we would need to evaluate how reliably different coders generate the same graph and, if they differ, by how much. That is, we would need to test the inter-coder reliability.

Once the Normative Graphs are in place, we still need to determine what the permitted actions in each case are. Here, we think, domain experts are the best access we have to any moral truth.²⁹ It would be especially important to see if experts (or even a single expert) produce consistent results. However, as both challenges need considerable (human) resources to tackle, they are left for future work. Instead of human coders, it might also be possible to extract the normative structure using software alone. Here, the task is to use further machine learning tools to infer the graph structure, either entirely from the description of cases (in addition to some training material) or from semi-structured *normative data*.³⁰

Lastly, we believe that our framework is flexible enough to be applied to the same domain as both GenEth and Freedman et al.’s framework: medical ethics and the dataset of the *Moral Machine* experiments (see (Awad et al., 2020b, 2018) and (Awad et al., 2020a) for a discussion of GenEth in the context of this dataset). However, it is not always trivial to directly translate the normative structure and datasets that these frameworks rely on to our structure. Still, we are optimistic that our framework can achieve better results here.

So let us briefly recapitulate the main points of this chapter. We have proposed to use the formal version of the kitchen scale model of reasons that we introduced in Chapter 3 as a starting point to develop a hybrid machine ethics framework. We used a set of cases as

²⁹This is a somewhat controversial opinion. In the field of machine ethics, this is also the favoured approach of (Anderson and Anderson, 2018). Other approaches suggest taking a more democratic stance and let the public decide, this is the approach, at least implicitly, behind the *Moral Machine* (see (Awad et al., 2020b, 2018)). There are also more nuanced approaches for public and expert participation suggested, as in (Sinnott-Armstrong et al., 2021). We do not mean to take a strong stance here on which of these will ultimately turn out to be the best.

³⁰We think there is an interesting connection here to the problem of causal (or probabilistic) inference. Once a causal graph in the style of (Pearl, 2009) or a simple Bayesian network is in place, giving appropriate weights to the nodes in the graph is a sometimes difficult but often tractable problem. The question of which causal graph (or Bayesian network) is a good fit for modeling some phenomenon is an entirely different (and, arguably, much more difficult) question. We see the generation of Normative Graphs as similarly challenging.

training data that allowed us to estimate the weights of normative considerations using a genetic algorithm. We then used these weights to determine the morally correct action(s) in novel cases that are composed of (some) of the same considerations. We then tested the framework on synthetic data. We showed that the prediction accuracy remains quite high even when the set of considerations (and thus the set of possible cases) becomes quite large. Furthermore, we showed that the choice of the function determining the final weights of considerations does not play an important role and all functions (except perhaps the rather simplistic Additive Function) perform well, even when the synthetic training data is generated by a different function. We also showed that noisy training data does not impact the prediction accuracy much. Lastly, we showed that roughly 60–70 percentage points of accuracy, that is, the majority of normative information, is stored in the graph structure and not in the weights.

We then compared our model to related work by (Freedman et al., 2020), (Anderson and Anderson, 2018), and (Canavotto and Horty, 2022) and argued that our model can overcome some limitations found in the frameworks proposed by them. Specifically, we believe, our model is more philosophically justified and requires much less training data to work.

5. Conclusion

5.1. Revisiting the Idea of Formal Ethics

We started this thesis by talking about the idea of *formal ethics*. It is fitting to return to this topic here. In Chapter 1, we said that this thesis will mostly deal with the question of modeling how reasons relate to oughts (and requirements) formally. But now that we have seen various ways to answer this question, we can spell out what formal ethics can do (and what we have done) in much more detail.

In effect, what we have seen are three ways to approach formal ethics, one in each main chapter of this thesis. Next, we will look at what role the *formal* part of formal ethics played.

In Chapter 2, we used logical tools to study the underexplored phenomenon of supererogation (and requiring and justifying reasons). There, we used formal tools that allowed us to stay precise to solve philosophically interesting paradoxes. But nothing about this problem *required* formal tools to be solved. It is entirely plausible that there would have been a way to resolve the paradoxes with just the standard conceptual tools of philosophy. Why then bother with formal tools at all? The answer is, we claim, that sometimes formal tools are the best tools to get the job done. In our case, we could make the paradoxes precise and show how to resolve them in a unified framework that makes our assumptions very explicit. On this approach, formal ethics is just ethics with formal tools. Sometimes, these tools are the best tools for the job, and sometimes they are not.

In Chapter 3, the approach was different. Here, our goal was to develop a formal version of the kitchen scale model by building it up from the fundamentals. Large parts of this chapter were predominantly exploratory. We identified the building blocks we needed for the model to succeed: the structure of Normative Graphs, a function that determines the final weights of considerations and a function that determines the deontic statuses of actions based on those final weights. However, as there is no consensus as to what these three building blocks look like, we undertook a principle-based analysis. We could only do so because we were operating in a formal setting. It would have been impossible to do this without a formal framework, much like it is impossible to discuss the laws of motion without a mathematical framework to predict experimental outcomes. So, on

5. Conclusion

this approach, formal ethics is not just a tool, but a methodology of getting ethics right.

Lastly, in Chapter 4, we looked at formal ethics through the lens of *computational ethics*.¹ Here, unlike in Chapter 3, we take the model as given and explore applications with it. What the formal element then allows us is to make ethics machine-implementable. That means that on this approach, what formal ethics allows us to do is build applications on top of ethical theories. Something that could not have been done with their non-formal analogues.

So, we have seen three different takes on what formal ethics brings to the table. We believe that we have shown that this is more than just *precision* or *proofs*.

5.2. Summing Up

So let us briefly summarize the thesis, its main steps and results. In Chapter 1, we briefly introduced the concept of normative reasons. We then introduced so-called dual role views of reasons that distinguish between justifying and requiring reasons. Next, we discussed the *kitchen scale* model of reasons. The last part of Chapter 1 was spent on discussing existing formal models of reasons.

In Chapter 2, we discussed paradoxes surrounding supererogation. We first introduced a simple model of reasons and showed that while it is a good fit for modeling simple cases, it could not adequately model certain cases involving supererogation. We then supplemented the model so that it could accommodate both justifying and requiring reasons and express a distinction between two deontic modalities, requirements and oughts. This updated model then allowed us to formalize the paradoxes and resolve them. We did this for each paradox by showing that an intuitive principle that the paradoxes rely on is not valid in our model. We were also able to show why they seemed so intractable at first, by showing that principles close to the ones we argued should be given up *do* hold in our model. We then showed that for exactly one of the paradoxes, the *Intransitivity Paradox*, the distinction between justifying and requiring reasons is not actually needed, as long as one accepts a certain way of how reasons determine what one is permitted to do. We argued that our approach is in some respects quite similar to recent non-formal approaches in the literature, but different in others. On the formal side, we argued that competing accounts of supererogation fare worse than our approach. Lastly, we showed that there is a perhaps surprising overlap between our formal model and formal models of explicit permission.

¹We use the term *computational ethics* instead of machine ethics here to focus on the way ethics is *computed*. We believe that for present purposes, at least, machine ethics and computational ethics can be used interchangeably.

In Chapter 3, our goal was to develop a formal version of the kitchen scale model that, roughly, says that reasons come with weights and what one ought to (or is required to) do is determined by the joint weight of reasons for and against. We argued that for such a model to be viable it must be a *holist* model, that is the weight of a reason must be able to change depending on the context it is found in. We then argued that the most plausible version of holism is what we have called *constrained holism*, that says that this context-dependent weight of reasons is generated through the interaction of a default weight that reasons possess and a select class of normative considerations that are not reasons: modifiers and undercutters.

Next, we made this idea formal through the introduction of *Normative Graphs*. We argued that the kitchen scale model consists of two additional modules. One module that governs the final weight of reasons and one module that weighs the final weight of reasons to determine which actions are permitted in a given situation. In order to analyze these two modules, we performed a principle-based analysis of possible semantics and introduced several candidates.

Next, we applied the formal model to philosophical issues. We first addressed challenges to the *weighing* and then to the *weighting* of reasons. We argued that the choice of weighing functions is not arbitrary and that challenges to aggregation can be answered within the model. We argued that our model provides an answer to what the weight of a reason *means*. This led us to show that in many cases we can *reverse engineer* the weights of reasons. We also responded to a criticism that we cannot make sense of the weight of reasons at all, by showing that they arise naturally out of the *weighing* of reasons and can be learned through reverse engineering.

We then extended the model to account for the difference between requiring and justifying reasons. We argued that in order to incorporate this distinction, we need to change the way reasons are *weighed* because justifying reasons, unlike requiring reasons, do not cancel out. Lastly, we compared our model to related formal and non-formal models.

In Chapter 4, we applied the model from the previous chapter to create a machine ethics framework. We discussed how we can learn normative information from a set of cases where we know which actions are morally correct. We then made this idea precise by introducing a genetic algorithm that allows us to learn the weights of considerations from a set of cases. Next, we presented a mechanism to evaluate the framework. We ran several experiments to determine the predictive accuracy of our model and found that it performs well even if the training data is sparse or noisy. We also discovered that the majority of information in our model is not stored in the weights of considerations, but in the graph structure of a Normative Graph. Finally, we compared our model to related hybrid approaches to machine ethics and highlighted what we took to be the advantages of our approach.

A. Appendix

A.1. Updated Definitions for Chapter 2

Definition A.1.1 (Brewka scenarios for Totally Ordered Contexts).

Let $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$ be a context where $<$ totally orders $\mathcal{J}$. Then $\mathcal{B}$ is the Brewka scenario of Δ just in case $\mathcal{B} = \bigcup_{i \geq 0} \mathcal{B}_i$, where the sequence $\mathcal{B}_0, \mathcal{B}_1, \mathcal{B}_2, \dots$ is defined as follows:

$$\begin{aligned} \mathcal{B}_0 &= \emptyset, \\ \mathcal{B}_{i+1} &= \begin{cases} \mathcal{B}_i & \text{if } \text{Active}_\Delta(\mathcal{B}_i) = \emptyset \\ \mathcal{B}_i \cup \text{Max}_{<}(\text{Active}_\Delta(\mathcal{B}_i)) & \text{otherwise} \end{cases} \end{aligned}$$

Definition A.1.2 (Brewka Scenarios). Let $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$ be any context. Then $\mathcal{B}$ is a Brewka scenario based on Δ just in case $\mathcal{B}$ is the Brewka scenario of some context $(\mathcal{W}, \mathcal{R}, \mathcal{J}, <')$ where $<'$ is a total order extending $<$.

A.2. Proofs for Chapter 2

Lemma A.2.1. For any context Δ and any restricted context $\Delta^{\mathcal{D}}$ based on it, there is always a Brewka scenario based on $\Delta^{\mathcal{D}}$.

Proof. This follows from the construction of the Brewka Scenarios. Pick any order that extends the order of $\Delta^{\mathcal{D}}$ to a total order (such an extension exists for every partial order). Then, note that in the construction of $\mathcal{B}$, the $\mathcal{B}_i$ are $\subseteq$ -ordered, that is, each $\mathcal{B}_i$ contains all its predecessors and if $\mathcal{B}_j = \mathcal{B}_{j+1}$, then for all $\mathcal{B}_k$ with $k > j$, $\mathcal{B}_k = \mathcal{B}_j$. Since Δ is finite, there is a maximal element $\mathcal{B}_n$, such that for every $m > n$, $\mathcal{B}_m = \mathcal{B}_n$; possibly, this n is 0. By the construction of $\mathcal{B}$, $\mathcal{B}_n$ is a Brewka scenario. As we can always construct one explicitly in this way, every $\Delta^{\mathcal{D}}$ has one. $\square$

Observation 2.3.1. For any context $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$,

A. Appendix

- if $\Delta \vdash Must(X)$, then $\Delta \vdash Ought(X)$; and
- if $\Delta \vdash Ought(X)$, then $\Delta \vdash Can(X)$.

Proof. We prove both statements in turn. Assume first that $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <) \vdash Must(X)$. Then, by definition of $Must(\cdot)$, for every $\mathcal{D}$, such that $\mathcal{R} \subseteq \mathcal{D} \subseteq \mathcal{J}$, and every Brewka scenario $\mathcal{B}$ based on $\Delta^{\mathcal{D}}$, $\mathcal{W} \cup Conclusion(\mathcal{B}) \vdash X$. Since, by construction of a context, $\mathcal{R} \subseteq \mathcal{R} \subseteq \mathcal{J}$, we get that for every Brewka scenario $\mathcal{B}^*$ based on $\Delta^{\mathcal{R}}$, $\mathcal{W} \cup Conclusion(\mathcal{B}^*) \vdash X$, which means, by definition, that $\Delta \vdash Ought(X)$.

Next, assume that $\Delta \vdash Ought(X)$, by definition, together with Lemma A.2.1, this implies that there is some Brewka scenario $\mathcal{B}$ of $\Delta^{\mathcal{R}}$, such that $\mathcal{W} \cup Conclusion(\mathcal{B}) \vdash X$, and hence $\Delta \vdash Can(X)$. $\square$

Lemma A.2.2. *For any context $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$, the conclusions of any Brewka scenario $\mathcal{B}$ of Δ are consistent with $\mathcal{W}$.*

Proof. First, note that $\mathcal{W}$ is a consistent set by definition. Let $<'$ be an arbitrary extension of $<$ to a total order. We then show that for every step of the construction of the Brewka scenario $\mathcal{B}$ based on the context extended to this total order, the conclusions of $\mathcal{B}_i$ are still consistent with $\mathcal{W}$. As $\mathcal{W}$ is consistent, (the conclusion of) $\emptyset$ are consistent with $\mathcal{W}$. By the definition of an active reason, each step preserves consistency and so each $\mathcal{B}_i$ is consistent with $\mathcal{W}$.

It is left to show that the conclusions of the union $\mathcal{B} = \bigcup_{i \geq 0} \mathcal{B}_i$ are consistent with $\mathcal{W}$ as well. First note, that the $\mathcal{B}_i$ are $\subseteq$ -ordered, that is, each $\mathcal{B}_i$ contains all its predecessors. Next, assume that the conclusions of $\mathcal{B}$ are not consistent with $\mathcal{W}$. This means that there is a *finite* number of elements in $\mathcal{B}$ whose conclusions are jointly inconsistent with $\mathcal{W}$. But then, there is some n , such that $\mathcal{B}_n$ contains these elements. This implies that the conclusions of $\mathcal{B}_n$ are inconsistent with $\mathcal{W}$ and this is a contradiction as we have shown all $\mathcal{B}_i$ have conclusions consistent with $\mathcal{W}$. $\square$

Observation 2.3.2. *For any context $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$, none of the following hold:*

- $\Delta \vdash Ought(\perp)$
- $\Delta \vdash Must(\perp)$
- $\Delta \vdash Can(\perp)$

Proof. It is enough to show that $\Delta \vdash \text{Can}(\perp)$ does not hold, since by Proposition 2.3.1, if one of the other statements was to hold, so would $\Delta \vdash \text{Can}(\perp)$. But by Lemma A.2.2, every Brewka scenario is consistent with $\mathcal{W}$ and hence it is not the case that $\Delta \vdash \text{Can}(\perp)$. $\square$

Proposition 2.3.4. *Let $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$ be a dual role context and $\Delta' = (\mathcal{W}, \mathcal{R}, <')$ be a single role context (such that they have the same $\mathcal{W}$ and $\mathcal{R}$) with $<'$ being the restriction of $<$ to $\mathcal{R}$. Then:*

- $\Delta \vdash \text{Ought}(X)$ iff $\Delta' \vdash \text{Ought}^*(X)$

and additionally, if $\mathcal{J} = \mathcal{R}$, we get

- $\Delta \vdash \text{Must}(X)$ iff $\Delta \vdash \text{Ought}(X)$ iff $\Delta' \vdash \text{Ought}^*(X)$
- $\Delta \vdash \text{Can}(X)$ iff $\Delta' \vdash \text{Can}^*(X)$ iff there is a Brewka scenario $\mathcal{B}$ based on $\Delta^{\mathcal{R}}$ or on Δ or on Δ' such that $\text{Conclusion}(\mathcal{B}) \vdash X$.

Proof. Since $\Delta^{\mathcal{R}} = \mathcal{R}$, the fact that $\Delta \vdash \text{Ought}(X)$ iff $\Delta' \vdash \text{Ought}^*(X)$ follows straight from their respective definitions.

For the second claim, note that there is but a single restricted context of Δ , namely $\Delta = \Delta^{\mathcal{R}}$ itself and so the definitions of $\text{Must}(\cdot)$ and $\text{Ought}(\cdot)$ coincide for this context. The second half of the statement follows from the first claim.

The third claim likewise just follows from the fact that there is only a single restricted context and the definitions. $\square$

Observation 2.3.3. *Let $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$ be an arbitrary context. Then*

- $\Delta \vdash \text{Ought}(X \wedge Y)$ just in case, both $\Delta \vdash \text{Ought}(X)$ and $\Delta \vdash \text{Ought}(Y)$ and
- $\Delta \vdash \text{Must}(X \wedge Y)$ just in case, both $\Delta \vdash \text{Must}(X)$ and $\Delta \vdash \text{Must}(Y)$.

Proof. The proofs follow straight from the definitions and propositional logic. $\square$

Lemma A.2.3. *Let $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <')$ be a context with $<'$ a total order on $\mathcal{J}$, $\mathcal{B}$ the Brewka scenario of Δ with $\mathcal{W} \cup \text{Conclusion}(\mathcal{B}) \vdash X$, and $\mathcal{B}^*$ the Brewka scenario of the context $\Delta[X \vee Y] = (\mathcal{W} \cup \{X \vee Y\}, \mathcal{R}, \mathcal{J}, <')$. Then $\mathcal{B} = \mathcal{B}^*$.*

A. Appendix

Proof. Before getting into the proof, note that both $\mathcal{B}$ and the sequence of elements it is constructed out of are unique. This is because at each step i of the construction of $\mathcal{B}$ there is at most one $<'$ -maximal reason in $Active_{\Delta}(\mathcal{B}_i)$. We will show that, for each step i , $\mathcal{B}_i = \mathcal{B}_i^*$. We do this by induction on i .

The base case is trivial: $\mathcal{B}_0 = \emptyset = \mathcal{B}_0^*$.

For the induction step, assume that $\mathcal{B}_i = \mathcal{B}_i^*$. Given our definition of Brewka scenarios, it's enough to establish that $Max_{<'}(Active_{\Delta}(\mathcal{B}_i)) = Max_{<'}(Active_{\Delta[X \vee Y]}(\mathcal{B}_i^*))$. So that's what we turn to.

$\subseteq$: Consider some $r \in Max_{<'}(Active_{\Delta}(\mathcal{B}_i))$. Then $r \in \mathcal{B}_{i+1} \subseteq \mathcal{B}$. We know that $\mathcal{W} \cup Conclusion(\mathcal{B}) \vdash X$, and hence that $\mathcal{W} \cup Conclusion(\mathcal{B}) \vdash X \vee Y$. By Lemma A.2.2, $\mathcal{W} \cup Conclusion(\mathcal{B})$ is consistent. Hence, as it entails $X \vee Y$, it is also consistent with $X \vee Y$. Since $\mathcal{W} \cup Conclusion(r) \cup Conclusion(\mathcal{B}_i)$ is a subset of $\mathcal{W} \cup Conclusion(\mathcal{B})$, we can be sure that $\mathcal{W} \cup \{X \vee Y\} \cup Conclusion(\mathcal{B}_i) \cup Conclusion(r) \not\vdash \perp$. This suffices to conclude that $r \in Active_{\Delta[X \vee Y]}(\mathcal{B}_i)$. We still need to show that r is $<'$ -maximal in this set. We know that r is $<'$ -maximal in $Active_{\Delta}(\mathcal{B}_i)$. Hence, for every $r' > r$ such that $r' \notin \mathcal{B}_i$, $r' \notin Active_{\Delta}(\mathcal{B}_i)$. But this means that, for every such r' , $\mathcal{W} \cup Conclusion(\mathcal{B}_i) \cup Conclusion(r') \vdash \perp$, and, hence, by monotonicity, that $\mathcal{W} \cup \{X \vee Y\} \cup Conclusion(\mathcal{B}_i) \cup Conclusion(r') \vdash \perp$. This means that each such r' is not in $Active_{\Delta[X \vee Y]}(\mathcal{B}_i)$, and that r is indeed maximal here. By the inductive hypothesis, $\mathcal{B}_i = \mathcal{B}_i^*$, and so we have shown that $r \in Max_{<'}(Active_{\Delta[X \vee Y]}(\mathcal{B}_i^*))$.

$\supseteq$: Suppose that there is an $r \in Max_{<'}(Active_{\Delta[X \vee Y]}(\mathcal{B}_i^*))$. Further, suppose, toward a contradiction, that $r \notin Max_{<'}(Active_{\Delta}(\mathcal{B}_i))$. Either $r \in Active_{\Delta}(\mathcal{B}_i)$ or not. If not, then either (i) $r \in \mathcal{B}_i$, or (ii) $\mathcal{W} \cup Conclusion(\mathcal{B}_i) \cup Conclusion(r) \vdash \perp$. If (i), then $\mathcal{B}_i \neq \mathcal{B}_i^*$ contradicting our assumption. If (ii), then $\mathcal{W} \cup Conclusion(\mathcal{B}_i^*) \cup Conclusion(r) \vdash \perp$, and $r \notin Active_{\Delta[X \vee Y]}(\mathcal{B}_i^*)$ after all. Hence, $r \in Active_{\Delta}(\mathcal{B}_i)$, but r is not $<'$ -maximal. Let r' be the $<'$ -maximal reason in $Active_{\Delta}(\mathcal{B}_i)$. Now we can reuse the argument we made use of above to conclude that $r' \in Active_{\Delta[X \vee Y]}(\mathcal{B}_i^*)$, and that r is not the maximal element in $(Active_{\Delta[X \vee Y]}(\mathcal{B}_i^*))$ after all. This gives us a contradiction. $\square$

Observation 2.4.1. For any context Δ ,

- if $\Delta \vdash Can(X)$ and $\Delta \vdash Must(\neg Y)$, then $\Delta \vdash Can(X|X \vee Y)$;
- if $\Delta \vdash Ought(X)$ and $\Delta \vdash Must(\neg Y)$, then $\Delta \vdash Ought(X|X \vee Y)$.

Proof. We establish claim (i) by proving a stronger claim, namely, that if $\Delta \vdash Can(X)$, then $\Delta \vdash Can(X|X \vee Y)$. Suppose that $\Delta \vdash Can(X)$. It follows that there is a stable scenario $\mathcal{D}$ based on Δ and a total order $<'$ on $\mathcal{D}$ that extends $<$ such that $\mathcal{W} \cup$

$\text{Conclusion}(\mathcal{B}) \vdash X$ for the Brewka scenario based on $(\mathcal{W}, \mathcal{R}, \mathcal{D}, <')$. Note that $\mathcal{D}$ is a stable scenario of $\Delta[X \vee Y]$, and that $<'$ is a total order extending $<$ in this restricted conditional context as well. Set Δ^* to be the context $(\mathcal{W} \cup \{X \vee Y\}, \mathcal{R}, \mathcal{D}, <')$. By Lemma A.2.3, we know that $\mathcal{B}^* = \mathcal{B}$ where $\mathcal{B}^*$ is the Brewka scenario based on Δ^* . Since $\mathcal{W} \cup \text{Conclusion}(\mathcal{B}) \vdash X$, we immediately get $\mathcal{W} \cup \text{Conclusion}(\mathcal{B}^*) \vdash X$, and, by monotonicity of classical logic, $\mathcal{W} \cup \{X \vee Y\} \cup \text{Conclusion}(\mathcal{B}^*) \vdash X$. This means that there is a stable scenario of $\Delta[X \vee Y]$, namely, $\mathcal{D}$, and a Brewka scenario based on $\Delta[X \vee Y]^{\mathcal{D}}$, namely, $\mathcal{B}^*$, such that $\mathcal{W} \cup \{X \vee Y\} \cup \text{Conclusion}(\mathcal{B}^*) \vdash X$. Given our definition of permissions, this is enough to conclude that $\Delta[X \vee Y] \sim \text{Can}(X)$, and hence that $\Delta \sim \text{Can}(X|X \vee Y)$.

For Claim (ii), we prove something stronger, namely, that if $\Delta \sim \text{Ought}(X)$, then also $\Delta \sim \text{Ought}(X|X \vee Y)$. Suppose that $\Delta \sim \text{Ought}(X)$. This means that, for any Brewka scenario $\mathcal{B}$ of the context $\Delta^{\mathcal{R}}$, we have $\mathcal{W} \cup \text{Conclusion}(\mathcal{B}) \vdash X$. What we need to show is that, for any Brewka scenario $\mathcal{B}^*$ based on $\Delta[X \vee Y]^{\mathcal{R}}$, we have $\mathcal{W} \cup \{X \vee Y\} \cup \text{Conclusion}(\mathcal{B}^*) \vdash X$. Suppose, toward a contradiction, that this wasn't the case. So there is a context $\Delta^* = (\mathcal{W} \cup \{X \vee Y\}, \mathcal{R}, \mathcal{R}, <')$, where $<'$ extends $<$ to a total order over $\mathcal{R}$, such that $\mathcal{W} \cup \{X \vee Y\} \cup \text{Conclusion}(\mathcal{B}^*) \not\vdash X$ for the Brewka scenario $\mathcal{B}^*$ based on Δ^* . Consider the context $(\mathcal{W}, \mathcal{R}, \mathcal{R}, <')$. From above, we can be sure that, for the Brewka scenario $\mathcal{B}$ based on it, we have $\mathcal{W} \cup \text{Conclusion}(\mathcal{B}) \vdash X$. By Lemma A.2.3, we have $\mathcal{B}^* = \mathcal{B}$. (Recall that $\mathcal{B}^*$ is unique.) Hence, $\mathcal{W} \cup \text{Conclusion}(\mathcal{B}^*) \vdash X$, and, by the monotonicity of classical logic, $\mathcal{W} \cup \{X \vee Y\} \cup \text{Conclusion}(\mathcal{B}^*) \vdash X$. And this is a contradiction. $\square$

Lemma A.2.4. *Given a context $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$ where $<$ is a total order, let $\mathcal{B}^1$ be the Brewka scenario of $\Delta[X \vee Y]$, $\mathcal{B}^2$ be the Brewka scenario of $\Delta[Y \vee Z]$ and $\mathcal{B}^3$ that of $\Delta[X \vee Z]$.*

Then, if $\mathcal{W} \cup \{X \vee Y\} \cup \text{Conclusion}(\mathcal{B}^1) \vdash X$ and $\mathcal{W} \cup \{Y \vee Z\} \cup \text{Conclusion}(\mathcal{B}^2) \vdash Y$, then $\mathcal{W} \cup \{X \vee Z\} \cup \text{Conclusion}(\mathcal{B}^3) \vdash X$.

Proof. We will prove this by contradiction. Assume that $\mathcal{W} \cup \{X \vee Y\} \cup \text{Conclusion}(\mathcal{B}^1) \vdash X$ and $\mathcal{W} \cup \{Y \vee Z\} \cup \text{Conclusion}(\mathcal{B}^2) \vdash Y$, but $\mathcal{W} \cup \{X \vee Z\} \cup \text{Conclusion}(\mathcal{B}^3) \not\vdash X$.

To do this, we will show that, given this assumption, $\mathcal{B}^1 = \mathcal{B}^2 = \mathcal{B}^3$. Then, by assumption, $\mathcal{W} \cup \{X \vee Y\} \cup \text{Conclusion}(\mathcal{B}^3) \vdash X$ and $\mathcal{W} \cup \{Y \vee Z\} \cup \text{Conclusion}(\mathcal{B}^3) \vdash Y$, and by the transitivity of classical consequence, so too $\mathcal{W} \cup \{X \vee Z\} \cup \text{Conclusion}(\mathcal{B}^3) \vdash X$ which contradicts our assumption.

All that is left is to show that $\mathcal{B}^1 = \mathcal{B}^2 = \mathcal{B}^3$.

By definition, the Brewka scenarios are constructed iteratively, thus let $\mathcal{B}^1 = \bigcup_{i \geq 0} \mathcal{B}_i^1$,

A. Appendix

$\mathcal{B}^2 = \bigcup_{i \geq 0} \mathcal{B}_i^2$, $\mathcal{B}^3 = \bigcup_{i \geq 0} \mathcal{B}_i^3$. Note that for each Brewka scenario, each element in the union is uniquely determined by the definition.

Assume for contradiction that it is not the case that $\mathcal{B}^1 = \mathcal{B}^2 = \mathcal{B}^3$. Then, there is a largest $n > 0$, such that $\mathcal{B}_n^1 = \mathcal{B}_n^2 = \mathcal{B}_n^3$, we will simply refer to this set as $\mathcal{B}_n$. As n is the largest such integer, we know that it is not the case that $Active_{\Delta[X \vee Y] \mathcal{R}}(\mathcal{B}_n) = Active_{\Delta[Y \vee Z] \mathcal{R}}(\mathcal{B}_n) = Active_{\Delta[X \vee Z] \mathcal{R}}(\mathcal{B}_n)$. As there is at most one reason in each of these sets, there is a reason in one of the sets that is $<$ -maximal with regard to all of these sets. Call that reason r .

Next we prove that in whichever set this reason lies, we reach a contradiction. This in turn implies that the maximal reason in each of the set is the same and hence $\mathcal{B}^1 = \mathcal{B}^2 = \mathcal{B}^3$, proving our theorem.

Case 1: Assume that $r \in Active_{\Delta[X \vee Y]}(\mathcal{B}_n)$. As r cannot be in the other two sets (else they would be equal) this gives us three possibilities which we will consider in turn.

Case 1.1: $r \notin Active_{\Delta[Y \vee Z]}(\mathcal{B}_n)$ but $r \in Active_{\Delta[X \vee Z]}(\mathcal{B}_n)$.

Case 1.2: $r \in Active_{\Delta[Y \vee Z]}(\mathcal{B}_n)$ but $r \notin Active_{\Delta[X \vee Z]}(\mathcal{B}_n)$.

Case 1.3: $r \notin Active_{\Delta[Y \vee Z]}(\mathcal{B}_n)$ and $r \notin Active_{\Delta[X \vee Z]}(\mathcal{B}_n)$.

First, consider case 1.1. As $r \notin Active_{\Delta[Y \vee Z]}(\mathcal{B}_n)$, $\mathcal{W} \cup Conclusion(r) \cup Conclusion(\mathcal{B}_n) \vdash \neg Y \wedge \neg Z$, by definition. As r is an element of $Active_{\Delta[X \vee Z]}(\mathcal{B}_n)$ (and is a maximal reason in this set), we get that $r \cup \mathcal{B}_n \subseteq \mathcal{B}^3$. But then, $\mathcal{W} \cup Conclusion(\mathcal{B}^3) \vdash \neg Z$ by monotonicity of classical consequence and so $Conclusion(\mathcal{B}^3) \cup \mathcal{W} \cup \{X \vee Z\} \vdash X$. This contradicts our assumption.

For both case 1.2 and 1.3, the sought after contradiction is easy to reach. In both cases, as $r \notin Active_{\Delta[X \vee Z]}(\mathcal{B}_n)$, $\mathcal{W} \cup Conclusion(r) \cup Conclusion(\mathcal{B}_n) \vdash \neg X \wedge \neg Z$. By the same argument as above, $r \cup \mathcal{B}_n \subseteq \mathcal{B}^1$ and hence $\mathcal{W} \cup Conclusion(\mathcal{B}^1) \cup \{X \vee Y\} \vdash \neg X$. This contradicts our assumptions as by Lemma A.2.2, Brewka scenarios are consistent.

So if $r \in Active_{\Delta[X \vee Y]}(\mathcal{B}_n)$, we reach a contradiction. The next possibility of the membership of r is given by case 2.

Case 2: Assume $r \in Active_{\Delta[Y \vee Z]}(\mathcal{B}_n)$. From Case 1, we know that $r \notin Active_{\Delta[X \vee Y]}$. This leaves us with two possibilities.

Case 2.1: $r \notin Active_{\Delta[X \vee Y]}(\mathcal{B}_n)$ but $r \in Active_{\Delta[X \vee Z]}(\mathcal{B}_n)$.

Case 2.2: $r \notin Active_{\Delta[X \vee Y]}(\mathcal{B}_n)$ and $r \notin Active_{\Delta[X \vee Z]}(\mathcal{B}_n)$.

The proof here follows the same layout as before. For both cases we get that $\mathcal{W} \cup \text{Conclusion}(r) \cup \text{Conclusion}(\mathcal{B}_n) \vdash \neg X \wedge \neg Y$ and, with the same argument as before as $r \cup \mathcal{B}_n \subseteq \mathcal{B}^2$, $\mathcal{W} \cup \text{Conclusion}(\mathcal{B}^2) \vdash \neg Y$, which contradicts our assumption.

This leaves one last case to consider. From Case 1 and Case 2, we know that r is neither in $\text{Active}_{\Delta[X \vee Y]}$ nor in $\text{Active}_{\Delta[Y \vee Z]}$. This only leaves one possible subcase.

Case 3: $r \in \text{Active}_{\Delta[X \vee Z]}(\mathcal{B}_n)$

Case 3.1: $r \notin \text{Active}_{\Delta[X \vee Y]}(\mathcal{B}_n)$ and $r \notin \text{Active}_{\Delta[Y \vee Z]}(\mathcal{B}_n)$.

For Case 3.1, we obtain $\mathcal{W} \cup \text{Conclusion}(r) \cup \text{Conclusion}(\mathcal{B}_n) \vdash \neg Y \wedge \neg Z$, so $\mathcal{W} \cup \text{Conclusion}(\mathcal{B}^3) \vdash \neg Z$ and hence $\mathcal{W} \cup \text{Conclusion}(\mathcal{B}^3) \cup \{X \vee Z\} \vdash X$. This, again, is a contradiction.

Thus, all possibilities result in a contradiction and hence $\mathcal{B}^1 = \mathcal{B}^2 = \mathcal{B}^3$. But then the fact that $\mathcal{W} \cup \{X \vee Z\} \cup \text{Conclusion}(\mathcal{B}^3) \vdash X$ follows from $\mathcal{W} \cup \{X \vee Y\} \cup \text{Conclusion}(\mathcal{B}^1) \vdash X$ and $\mathcal{W} \cup \{Y \vee Z\} \cup \text{Conclusion}(\mathcal{B}^2) \vdash Y$, by a simple application of propositional logic. This concludes the proof. $\square$

Proposition 2.4.2. *For any context Δ ,*

- *if $\Delta \vdash \text{Ought}(X|X \vee Y)$ and $\Delta \vdash \text{Ought}(Y|Y \vee Z)$ then $\Delta \vdash \text{Ought}(X|X \vee Z)$.*
- *if $\Delta \vdash \text{Must}(X|X \vee Y)$ and $\Delta \vdash \text{Must}(Y|Y \vee Z)$ then $\Delta \vdash \text{Must}(X|X \vee Z)$.*

We prove this in two steps. First we prove it for $\text{Must}(\cdot)$.

Proof. This proof makes ample use of Lemma A.2.4.

Let $\Delta = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <)$ and assume that $\Delta \vdash \text{Must}(X|X \vee Y)$ and $\Delta \vdash \text{Must}(Y|Y \vee Z)$.

Let $\mathcal{D}$ be an arbitrary set of reasons such that $\mathcal{R} \subseteq \mathcal{D} \subseteq \mathcal{J}$.

We want to show that for any total order extending the order of $\Delta[X \vee Z]^\mathcal{D}$, with $\mathcal{B}^3$ as its Brewka scenario, $\mathcal{W} \cup \{X \vee Z\} \cup \text{Conclusion}(\mathcal{B}^3) \vdash X$. Thus let $<'$ be an arbitrary total order extending the order of $\Delta[X \vee Z]^\mathcal{D}$. Note that thereby $<'$ is also a total order extending $<$. Let the context $\Delta' = (\mathcal{W}, \mathcal{R}, \mathcal{J}, <')$. Let $\mathcal{B}^1$ be the Brewka scenario of $\Delta'[X \vee Y]^\mathcal{D}$ and $\mathcal{B}^2$ the Brewka scenario of $\Delta'[Y \vee Z]^\mathcal{D}$. They are unique because $<'$ is a total order in the restricted context as well.

From the fact that $\text{Must}(X|X \vee Y)$ and $\text{Must}(Y|Y \vee Z)$ follow from Δ , by definition of

A. Appendix

$Must, \mathcal{W} \cup \{X \vee Y\} \cup Conclusion(\mathcal{B}^1) \vdash X$ and $\mathcal{W} \cup \{Y \vee Z\} \cup Conclusion(\mathcal{B}^2) \vdash Y$. By Lemma A.2.4, thus also $\mathcal{W} \cup \{X \vee Z\} \cup Conclusion(\mathcal{B}^3) \vdash X$.

As both $\mathcal{D}$ and $<'$ were chosen arbitrarily, it holds that for every restricted context $\Delta[X \vee Z]^{\mathcal{D}}$ and every Brewka scenario $\mathcal{B}$ based on it, $\mathcal{W} \cup \{X \vee Z\} \cup Conclusion(\mathcal{B}) \vdash X$ and thus, by definition, $\Delta \vdash Must(X|X \vee Z)$. $\square$

Next, we prove it for $Ought(\cdot)$.

Proof. The proof for this statement is the same as the previous proof when $\mathcal{D}$ is being held fixed as $\mathcal{R}$. $\square$

Observation 2.4.3. *Letting W abbreviate $\neg A \wedge \neg P$, We can see that*

- $\Delta_3^* \vdash Can^*(W|W \vee A)$
- $\Delta_3^* \vdash Can^*(A|A \vee P)$
- *but not $\Delta_3^* \vdash Can^*(W|W \vee P)$*
- *and thus not: If $Can^*(X|X \vee Y)$ and $Can^*(Y|Y \vee Z)$, then $Can^*(X|X \vee Z)$ for any X, Y, Z*

Proof. We prove each statement in turn. First, we prove that $\Delta_3^* \vdash Can^*(W|W \vee A)$. Note that there are two Brewka scenarios based on $\Delta_3^*[W \vee A]$, $\mathcal{B}_1 = \{!A\}$ and $\mathcal{B}_2 = \{\neg L\}$. From $Conclusion(\mathcal{B}_2) \cup \mathcal{W} \cup \{W \vee A\}$ we can derive W and hence $Can^*(W)$ follows from $\Delta_3^*[W \vee A]$ and thus $\Delta_3^* \vdash Can^*(W|W \vee A)$.

For the second statement: $\Delta_3^* \vdash Can^*(A|A \vee P)$, note that again, there are two Brewka Scenarios based on $\Delta_3^*[A \vee P]$: $\mathcal{B}_3 = \{!A\}$ and $\mathcal{B}_4 = \{!P, \neg L\}$. As we can derive A directly from $Conclusion(\mathcal{B}_3)$, we get that $\Delta_3^*[A \vee P] \vdash A$ and so $\Delta_3^* \vdash Can^*(A|A \vee P)$.

Next, we need to show that it is not the case that $\Delta_3^* \vdash Can^*(W|W \vee P)$. Here, we only have a single Brewka Scenarios of $\Delta_3^*[W \vee P]$, namely $\mathcal{B}_5 = \{!P, \neg L\}$. We know that $\mathcal{B}_5$ is consistent with $\mathcal{W} \cup \{W \vee P\}$, and so we know that $\neg P$ cannot be derived from $\mathcal{B}_5 \cup \mathcal{W} \cup \{W \vee P\}$. This in particular means that W cannot be derived and so $Can^*(W)$ does not follow from $\Delta_3^*[W \vee P]$ and so neither does $Can^*(W|W \vee P)$ follow from Δ_3^* . $\square$

Proposition 2.4.4. *For any single-type context Δ^* ,*

- if $\Delta^* \vdash \text{Ought}^*(X|X \vee Y)$ and $\Delta \vdash \text{Ought}^*(Y|Y \vee Z)$ then $\Delta \vdash \text{Ought}^*(X|X \vee Z)$.

Proof. The proof follows immediately from Proposition 2.3.4 together with Proposition 2.4.2.

□

A.3. Proofs for Chapter 3

Proofs for Table 3.1.

Proof. We will go through each principle and prove the statements for each function one-by-one.

First, Anonymity

- $f_1, \dots, f_4$ satisfy Anonymity: Follows immediately from the definitions of $f_1, \dots, f_4$.

Second, Unmodified Default

- $f_1, \dots, f_4$ satisfy Unmodified Default: Follows immediately from the definitions of $f_1, \dots, f_4$.

Third, Modeler's Delight

- $f_1, \dots, f_4$ satisfy Modeler's Delight: Follows immediately from the definitions of $f_1, \dots, f_4$, and u .

Fourth, No Effects from Spurious Modifiers

- $f_1, \dots, f_4$ satisfy No Effects from Spurious Modifiers: Proved inductively on the length of the chain from a consideration to the modifiers. For leaves (in Δ), the final weight is equal to the weight given by the weighting function, which is identical for both Normative Graphs. For the inductive step, either the modifiers in both graphs are the same (and have the same final weight) and so the final weight of the modified consideration is also identical or, there is some number of additional considerations that modify (or undercut) the consideration that all have a final weight of 0. In that case, the final weight remains the same by the definitions of $f_1, \dots, f_4$.

A. Appendix

Fifth, No Valence Flips for Modifiers

- $f_1, \dots, f_4$ satisfy No Valence Flips for Modifiers: Follows immediately from the definitions of $f_1, \dots, f_4$.

Sixth, Immediate Successors

- $f_1, \dots, f_4$ satisfy Immediate Successors: Follows immediately from the definitions of $f_1, \dots, f_4$, and u .

Seventh, Independence of Modifiers

- $f_1, \dots, f_4$ satisfy Independence of Modifiers: We can give the function s explicitly. For f_2 we define s as the product (of the weights), for the remaining functions as the sum (of the weights); then, the functions $f_1, \dots, f_4$ can be defined as usual.

Eighth, Weak Modifier Reciprocity

- $f_1, \dots, f_4$ satisfy Weak Modifier Reciprocity: Follows from the definitions given above and the definitions of $f_1, \dots, f_4$.

Ninth, Strong Modifier Reciprocity

- f_1 satisfies Strong Modifier Reciprocity: Follows from the fact that $(\sum_i^n a_i + y) - (\sum_j^m a_j + y) = \sum_i^n a_i - \sum_j^m a_j + (y - y) = \sum_i^n a_i - \sum_j^m a_j$.
- f_2 satisfies Strong Modifier Reciprocity: Follows from the fact that $\frac{\prod_i^n (1+a_i)^{(1+y)}}{\prod_j^m (1+a_j)^{(1+y)}} = \frac{\prod_i^n (1+a_i)^{1+y}}{\prod_j^m (1+a_j)^{1+y}} = \frac{\prod_i^n (1+a_i)}{\prod_j^m (1+a_j)}$.
- f_3 does not satisfy Strong Modifier Reciprocity: Consider as a counterexample a Normative Graph N with four considerations, a reason r with $w(r) = 1$ and two amplifiers of r , δ_1 and δ_2 and one attenuator of r , δ_3 with $w(\delta_3) = w(\delta_1) = 3$ and $w(\delta_2) = 5$. Then, $f_{3N}^w(r) = 1 * \frac{1+3+5}{1+3} = \frac{9}{4}$; but $f_{3N-\{\delta_1, \delta_3\}}^w(r) = 1 * \frac{1+5}{1} = 6$.
- f_4 satisfies Strong Modifier Reciprocity: Follows from the same fact as for f_1 .

Tenth, No Enablers

- f_1 does not satisfy No Enablers: As a counterexample, consider the Normative Graph N with one reason r and one amplifier δ of r such that $w(r) = 0$ and $w(\delta) = 100$. Then, $f_{1N}^w(r) = 100 \neq 0$.

- $f_2, \dots, f_4$ satisfy No Enablers: Follows immediately from the definitions of $f_2, \dots, f_4$.

Eleventh, Relativity – Final Weights

- f_1 satisfies Relativity – Final Weights: Follows, as the following statements are all equivalent: $k * f_N^w(\delta) = k * (w(\delta) + \sum_{a^+ \in A_\delta^+} f_{1N}^w(a^+) - \sum_{a^- \in A_\delta^-} f_{1N}^w(a^-)) = k * w(\delta) + \sum_{a^+ \in A_\delta^+} k * f_{1N}^w(a^+) - \sum_{a^- \in A_\delta^-} k * f_{1N}^w(a^-) = ((w * k)(\delta) + \sum_{a^+ \in A_\delta^+} f_{1N}^{w*k}(a^+) - \sum_{a^- \in A_\delta^-} f_{1N}^{w*k}(a^-)) = f_{1N}^{w*k}(\delta)$.
- $f_2, \dots, f_4$ do not satisfy Relativity – Final Weights: For a joint counterexample for each function, consider the Normative Graph N with a single reason r and a modifier δ of r such that $w(r) = 1$ and $w(\delta) = 3$. Then for all $f \in \{f_2, \dots, f_4\}$, $f_N^w(r) = 1 * (1 + 3) = 4$, but for $w * 4$ we get $f_N^{w*4}(r) = 4 * (1 + 12) = 52 \neq 16 = 4 * 4 = 4 * f_N^w(r)$.

Twelfth, Distinct Roles

- f_1 does not satisfy Distinct Roles: Trivial, as an undercutter of equal weight as some consideration reduces the weight to 0.
- $f_2, \dots, f_4$ satisfy Distinct Roles: Follows immediately from the definitions of $f_2, \dots, f_4$.

Thirteenth, Attenuators as Undercutters

- f_1 satisfies Attenuators as Undercutters: Let δ be some consideration in the Normative Graph N . Let x be the sum of $w(\delta)$ and the final weight of each amplifier of δ , let δ^* be an attenuator of δ that is not in N with $w^*(\delta^*) = x$ and otherwise equal to w ; then, by definition of f_1 , $f_{N+\{\delta^*\}}^{w^*} = 0$.
- $f_2, \dots, f_4$ do not satisfy Attenuators as Undercutters: Follows immediately from the fact that they satisfy Distinct Roles.

Fourteenth, F-Restriction

- f_1 satisfies F-Restriction for both $\mathbb{N}$ and $\mathbb{Q}$: This is simply because addition is a group operation on both the (additive) group of $\mathbb{N}$ and $\mathbb{Q}$.
- $f_2, \dots, f_4$ satisfies F-Restriction for $\mathbb{Q}$ but not for $\mathbb{N}$. The former is easy to see; for the latter, consider the case of a Normative Graph N with a single reason r and a single attenuator of r , δ with $w(r) = 1$ and $w(\delta) = 99$, then for all $f \in \{f_2, \dots, f_4\}$, $f_N^w(r) = \frac{1}{100} \notin \mathbb{N}$.

A. Appendix

Fifteenth, Atemporality

- f_1 does not satisfy Atemporality: Let $f = f_1$. Consider a reason r that is undercut by δ_u and amplified by δ_+ in some Normative Graph N . Let $w(r) = 5$, $w(\delta_u) = 5$ and $w(\delta_+) = 10$. Then, calculating first $f_{N-\delta_+}^w(r) = 0$ as it's undercut. Now, setting $w^*(r)$ to 0, we get $f_{N-\delta_u}^{w^*}(r) = 10$; But this is not equal to $f_N^w(r) = 0$.
- f_2 satisfies Atemporality. Let $f = f_2$; Let r be a reason in some Normative Graph N ; note that if r is undercut, the statement follows easily. If r is not undercut, let $\Delta^* \subset A_r^+ \cup A_r^- = \Delta'$. Then: $f_N^w(r) = w(r) * \frac{\prod_{a^+ \in A_r^+} 1+f_N^w(a^+)}{\prod_{a^- \in A_r^-} 1+f_N^w(a^-)} = w(r) * \frac{\prod_{a^+ \in A_r^+ \setminus \Delta^*} 1+f_N^w(a^+)}{\prod_{a^- \in A_r^- \setminus \Delta^*} 1+f_N^w(a^-)} * \frac{\prod_{a^+ \in A_r^+ \cap \Delta^*} 1+f_N^w(a^+)}{\prod_{a^- \in A_r^- \cap \Delta^*} 1+f_N^w(a^-)}$. As $w^*(r) = f_{N-\Delta^*}^w(r) = w(r) * \frac{\prod_{a^+ \in A_r^+ \setminus \Delta^*} 1+f_N^w(a^+)}{\prod_{a^- \in A_r^- \setminus \Delta^*} 1+f_N^w(a^-)}$, we get that this is equivalent to $f_{N-\Delta' \setminus \Delta^*}^{w^*}(r)$.
- f_3 and f_4 do not satisfy Atemporality: Let $f \in \{f_3, f_4\}$. Consider the Normative Graph N with a single reason r and two amplifiers of r , δ_1 , and δ_2 such that $w(r) = 1$ and $w(\delta_1) = w(\delta_2) = 3$. Then $f_{N-\delta_1}^w(r) = 1 * (1 + 3) = 4$, now setting $w^*(r) = 4$, we get that $f_{N-\delta_2}^{w^*}(r) = 4 * (1 + 4) = 20$. But this is not equal to $f_N^w(r) = 1 * (1 + (3 + 3)) = 7$.

Sixteenth, Modifier Scaling: Equality

- f_1 satisfies Modifier Scaling: Equality: Let $f = f_1$. Let $N = (\Delta, O)$ be some arbitrary Normative Graph, with n modifiers $\delta_1, \dots, \delta_n$ of some consideration r of the same role and weight. Let us designate that weight as x , let N' be the Normative Graph where these modifiers have been replaced by a single modifier δ of weight $n*x$ as described in the principle. Without loss of generality, assume that the modifiers are amplifiers. Then $f_N^w(r) = \max(0, \sum_{\delta' \in A_r^+ \setminus \{\delta_1, \dots, \delta_n\}} f_N^w(\delta') + \sum_{i=1}^n f_N^w(\delta_i) - \sum_{\delta' \in A_r^-} f_N^w(\delta')) = \max(0, \sum_{\delta' \in A_r^+ \setminus \{\delta_1, \dots, \delta_n\}} f_N^w(\delta') + n * f_N^w(\delta_i) - \sum_{\delta' \in A_r^-} f_N^w(\delta')) = \max(0, \sum_{\delta' \in A_r^+ \setminus \{\delta_1, \dots, \delta_n\}} f_{N'}^w(\delta') + f_{N'}^w(\delta_i) - \sum_{\delta' \in A_r^-} f_{N'}^w(\delta')) = f_{N'}^w(r)$.
- f_2 does not satisfy Modifier Scaling: Equality: Let $f = f_2$. As a counterexample, consider the Normative Graph N with one reason r and two amplifiers of r , δ_1, δ_2 , such that $w(r) = 1$ and $w(\delta_1) = w(\delta_2) = 3$ and the Normative Graph N' containing only r and a single amplifier δ with $w(\delta) = 2 * w(\delta_1) = 6$. Then $f_N^w(r) = 1 * (1 + 3)(1 + 3) = 16 \neq 7 = 1 * (1 + 6) = f_{N'}^w(r)$.
- f_3 and f_4 satisfy Modifier Scaling: Equality: Let $f \in \{f_3, f_4\}$. Let N be an arbitrary Normative Graph, r be some consideration of N . We will prove the statement for amplifiers, the proof for attenuators is analogous. Assume that

there are n amplifiers, $\delta_1, \dots, \delta_n$ of r with an identical weight of x . Let N' be the Normative Graph that results, if we replace these n amplifiers with a single amplifier δ of weight $n * x$ as described in the principle. Then, for $f = f_3$, $f_N^w(r) = w(r) * (\frac{1 + \sum_{a^+ \in A_r^+ \setminus \{\delta_1, \dots, \delta_n\}} f_N^w(a^+)}{1 + \sum_{a^- \in A_r^-} f_N^w(a^-)} + \frac{\sum_i^n f_N^w(\delta_i)}{1 + \sum_{a^- \in A_r^-} f_N^w(a^-)}) = w(r) * (\frac{1 + \sum_{a^+ \in A_r^+ \setminus \{\delta_1, \dots, \delta_n\}} f_N^w(a^+)}{1 + \sum_{a^- \in A_r^-} f_N^w(a^-)} + \frac{f_N^w(\delta)}{1 + \sum_{a^- \in A_r^-} f_N^w(a^-)}) = f_{N'}^w(r)$.

For $f = f_4$, assume without loss of generality that $\sum_{a^+ \in A_r^+} f_N^w(a^+) \geq \sum_{a^- \in A_r^-} f_N^w(a^-)$. Then, $f_N^w(r) = w(r) * (1 + \sum_{a^+ \in A_r^+} f_N^w(a^+) = w(r) * \sum_{a^+ \in A_r^+ \setminus \{\delta_1, \dots, \delta_n\}} f_N^w(a^+) + \sum_{i=1}^n f_N^w(\delta_i) - \sum_{a^- \in A_r^-} f_N^w(a^-) = w(r) + \sum_{a^+ \in A_r^+ \setminus \{\delta_1, \dots, \delta_n\}} f_N^w(a^+) + f_{N'}^w(\delta) - \sum_{a^- \in A_r^-} f_N^w(a^-) = f_{N'}^w(r)$.

Seventeenth, Modifier Scaling: Quality

- f_1 satisfies Modifier Scaling: Quality: Follows from the fact that f_1 satisfies Modifier Scaling: Equality.
- f_2 satisfies Modifier Scaling: Quality only for attenuators, not for amplifiers. The fact that it does satisfy it for attenuators follows from the fact that $\prod_i^n (1 + x) = (1 + x)^n \geq (1 + n * x)$. To see that it does not satisfy the principle for amplifiers, consider the Normative Graph N with a single reason r with $w(r) = 1$ and two amplifiers of r , δ_1 and δ_2 with $w(\delta_1) = w(\delta_2) = 3$ and the corresponding Normative Graph N' containing r and a single amplifier δ with $w(\delta) = 2 * w(\delta_1) = 6$. Then $f_N^w(r) = 1 * (1 + 3)(1 + 3) = 16 \not\leq 7 = 1 * (1 + 6) = f_{N'}^w(r)$.
- f_3 and f_4 satisfy Modifier Scaling: Quality: Follows from the fact that f_3 and f_4 satisfy Modifier Scaling: Equality.

Eighteenth, Modifier Scaling: Quantity

- f_1 satisfies Modifier Scaling: Quantity: Follows from the fact that f_1 satisfies Modifier Scaling: Equality.
- f_2 satisfies Modifier Scaling: Quantity only for amplifiers, not for attenuators. The fact that it does satisfy it for amplifiers follows from the fact that $\prod_i^n (1 + x) = (1 + x)^n \geq (1 + n * x)$. To see that it does not satisfy the principle for attenuators, consider the Normative Graph N with a single reason r with $w(r) = 1$ and two attenuators of r , δ_1 and δ_2 with $w(\delta_1) = w(\delta_2) = 3$ and the corresponding Normative Graph N' containing r and a single attenuator δ with $w(\delta) = 2 * w(\delta_1) = 6$. Then $f_N^w(r) = \frac{1}{(1+3)(1+3)} = \frac{1}{16} \not\geq \frac{1}{7} = \frac{1}{1+6} = f_{N'}^w(r)$.
- f_3 and f_4 satisfy Modifier Scaling: Quantity: Follows from the fact that f_3 and f_4 satisfy Modifier Scaling: Equality.

Proofs for Table 3.2.

Proof. We will go through each principle and prove the statements for each function one-by-one.

First, All Allowed

- g_1 satisfies All Allowed: Follows immediately from the definition of g_1 .
- g_2 does not satisfy All Allowed (for $k > 0$): Follows immediately from the definition of g_2 .
- $g_3, \dots, g_6$ satisfy All Allowed: Follows immediately from the definitions of $g_3, \dots, g_6$.

Second, Anonymity I

- $g_1, \dots, g_6$ satisfy Anonymity I: Follows immediately from the definitions of $g_1, \dots, g_6$.

Third, Anonymity II

- $g_1, \dots, g_6$ satisfy Anonymity II: Follows immediately from the definitions of $g_1, \dots, g_6$.

Fourth, No Effects from Spurious Reasons

- $g_1, \dots, g_6$ satisfy No Effects from Spurious Reasons: Follows immediately from the definitions of $g_1, \dots, g_6$.

Fifth, No Valence Flips for Reasons

- $g_1, \dots, g_6$ satisfy No Valence Flips for Reasons: Follows immediately from the definitions of $g_1, \dots, g_6$.

Sixth, Independence of Reasons

- $g_1, \dots, g_3$ satisfy Independence of Reasons: We can give the function s as sum of the (final) weights of the reasons for each option.
- g_4 and g_5 satisfy Independence of Reasons: We can give the function s explicitly as the sum of the (final) weights of the reasons for each option that are larger than the threshold l .

- g_6 satisfies Independence of Reasons: We can give the function s explicitly as the product of the (final) weights of the reasons for each option.

Seventh, Weak Reason Reciprocity

- $g_1, \dots, g_6$ satisfy Weak Reason Reciprocity: Follows immediately from the definitions of $g_1, \dots, g_6$ and the definitions of s above.

Eighth, Strong Reason Reciprocity

- $g_1, \dots, g_5$ satisfy Strong Reason Reciprocity: Follows immediately from the fact that $\sum_{a \in A} a + x - (\sum_{b \in B} b + x) = \sum_{a \in A} a - \sum_{b \in B} b + (x - x) = \sum_{a \in A} a - \sum_{b \in B} b$
- g_6 satisfies Strong Reason Reciprocity: Follows immediately from the fact that $\frac{\prod_{a \in A} a * x}{\prod_{b \in B} b * x} = \frac{\prod_{a \in A} x}{\prod_{b \in B} x} = \frac{\prod_{a \in A} a}{\prod_{b \in B} b}$.

Ninth, Relativity – Reasons

- $g_1, \dots, g_3$ satisfy Relativity – Reasons: Follows straightforwardly from the fact that $\sum_{a \in A} a \geq \sum_{b \in B} b$ iff $\sum_{a \in A} n * a \geq \sum_{b \in B} n * b$ for all $n > 0$.
- g_4 does not satisfy Relativity – Reasons: Let the threshold of g_4 be l . As a counterexample, consider the Normative Graph N with a single reason r in favour of o and two options o and o' with $0 < f_N^w(r) < l$; then $g_4(R_N) = \{o, o'\}$. now consider Normative Graph N' with a single reason r' and the same two options, such that $f_{N'}^w(r') = \frac{l}{f_N^w(r)} * l$ and so $f_{N'}^w(r') > l$ and hence $g_4(R_{N'}) = \{o\}$.
- g_5 does not satisfy Relativity – Reasons: Assume the same case as in the counterexample above, with the additional fact that the second threshold (k) is less than $f_{N'}^w(r)$.
- g_6 does not satisfy Relativity – Reasons: As a counterexample, consider a Normative Graph N with two options o and o' with two reasons r_1 and r_2 speaking against o with $f_N^w(r_1) = f_N^w(r_2) = 1$ and a single reason r' speaking against o' with $f_N^w(r') = 5$. Then $g_6(R_N) = \{o\}$. But for the Normative Graph N' with the same reasons, options and $f_{N'}^w(r_1) = f_{N'}^w(r_2) = 3 * f_N^w(r_1) = 3$ and $f_{N'}^w(r') = 3 * f_N^w(r') = 15$, then $g_6(R_{N'}) = \{o, o'\}$.

Tenth, No Dilemmas

- g_1 satisfies No Dilemmas: Follows immediately from the definition of g_1 .
- g_2 and g_3 do not satisfy No Dilemmas: As a counterexample, consider the Nor-

A. Appendix

mative Graph $N = (\Delta, O)$ with $O = \{o\}$ and $\Delta = \{r = (p, o, r^-)\}$, such that $w(r) > 0$. Then for $g = g_2$ or $g = g_3$, $g(R_N) = \emptyset$.

- g_4 satisfies No Dilemmas: Follows immediately from the definition of g_4 .
- g_5 does not satisfy No Dilemmas: We can use the same counterexample as for g_2 and g_3 , as long as $w(r) > l$.
- g_6 satisfies No Dilemmas: Follows immediately from the definition of g_6 .

Eleventh, Reasons Scaling: Quality

- $g_1, \dots, g_3$ satisfy Reasons Scaling: Quality: Follows immediately from the fact that $\sum_{a \in A} a + \sum_{i=1}^n b = \sum_{a \in A} a + n * b$
- g_4 does not satisfy Reasons Scaling: Quality for reasons against, but does satisfy it for reasons for. We first show that it does not satisfy it for reasons against. Let the threshold of g_4 be l . Consider the Normative Graph N_1 with two options o and o' and two reasons r_1 and r_2 speaking in against o such that $f_{N_1}^w(r_1) = f_{N_1}^w(r_2) = \frac{2l}{3} < l$. Next, consider the Normative Graph N'_1 , where we replace r_1 and r_2 with r_3 (also speaking in favour of o) where $f_{N'_1}^w(r_3) = \frac{4l}{3} > l$. Then, $g_4(R_{N_1}) = \{o, o'\}$, but $g_4(R_{N'_1}) = \{o'\}$ and so $o \notin g_4(R_{N_1})$. To show that the principle holds for reasons for, consider that $\sum_{r \in R_o^+(N)} f(r) \leq \sum_{r \in R_o^+(N - \{\delta_1, \dots, \delta_n\} + \delta^*)} f(r)$ where $\delta_1, \dots, \delta_n$ are reasons for o of the same strength and δ^* is a reason for o with n times that strength since if the weight of δ_1 is above the threshold l the sums will be equal and if not, the leftmost sum will be 0.
- g_5 does not satisfy Reasons Scaling: Quality for reasons against, but it does satisfy it for reasons for: Follows from the same kind of reasoning as for g_4 .
- g_6 does not satisfy Reasons Scaling: Quality for reasons for, but it does satisfy it for reasons against. The proof idea is the same as for the proof that function f_2 satisfies the Principle Modifier Scaling: Quality above.

Twelfth, Reasons Scaling: Quantity

- $g_1, \dots, g_3$ satisfy Reasons Scaling: Quantity: Proof is identical to Reasons Scaling: Quality.
- g_4 and g_5 do not satisfy Reasons Scaling: Quantity for reasons for, but they do satisfy it for reasons against: Proofs are the reverse of the proofs for Reasons Scaling: Quality.

- g_6 does not satisfy Reasons Scaling: Quantity for reasons against, but does satisfy it for reasons for. Proof is the reverse of the proof for Reasons Scaling: Quality.

Thirteenth, Difference Maker

- g_1 satisfies Difference Maker: Let r be an arbitrary reason for o of positive weight x and r' an arbitrary reason against o of positive weight y . Consider a Normative Graph N with two options o and o' and no considerations. then: $g_1(R_N) = \{o, o'\}$ but $g_1(R_{N+r}) = \{o\}$ and $g_1(R_{N+r'}) = \{o'\}$.
- g_2 satisfies Difference Maker: Let $k > 0$ be the threshold of g_2 ; let r be an arbitrary reason for o of positive weight x and r' an arbitrary reason against o of positive weight y . Consider a Normative Graph N with two options o and o' and a single consideration speaking in favour of o with weight $k - x$, then $g_2(R_N) = \{o, o'\}$ but $g_2(R_{N+r}) = \{o\}$. Analogously, consider N' with the same options and a single reason speaking in favour of o with weight k . Then $g_2(R_N) = \{o\}$ but $g_2(R_{N+r'}) = \{o, o'\}$.
- g_3 satisfies Difference Maker: Let r be an arbitrary reason for o of positive weight x and r' an arbitrary reason against o of positive weight y . Consider a Normative Graph N with two options o and o' and no considerations. then: $g_3(R_N) = \{o, o'\}$ but $g_3(R_{N+r}) = \{o\}$ and $g_3(R_{N+r'}) = \{o'\}$.
- g_4 and g_5 do not satisfy Difference Maker: Let $g \in \{g_4, g_5\}$ and let l be the threshold for the inclusion of reasons in g . Then, for any Normative Graph N and any reason r (for or against some option) with a weight less than k : $g(R_N) = g(R_{N+r})$.
- g_6 satisfies Difference Maker: same proof as for g_1 .

Fourteenth, Reasons Symmetry I

- g_1 satisfies Reasons Symmetry I: Follows immediately from the definition of g_1 and some algebraic manipulation.
- g_2 and g_3 do not satisfy Reasons Symmetry I: Let $g \in \{g_2, g_3\}$ Consider the Normative Graph N with two options o and o' and a single reason r against o (of positive weight larger than the threshold k of g_2). Then, $g(R_N) = \{o'\}$, but if we switch the action and polarity of r (in the Normative Graph N'), we get $g(R_{N'}) = \{o\}$.
- g_4 satisfies Reasons Symmetry I: Follows immediately from the definition of g_4 and some algebraic manipulation.
- g_5 does not satisfy Reasons Symmetry I: Same reasoning as for g_2 (given that all reasons are above the threshold l).

A. Appendix

- g_6 satisfies Reasons Symmetry I: Follows immediately from the definition of g_6 and some algebraic manipulation.

Fifteenth, Reasons Symmetry II

- g_1 satisfies Reasons Symmetry II: Same reasoning as for Reasons Symmetry I.
- g_2 and g_3 do not satisfy Reasons Symmetry II: Same reasoning as for Reasons Symmetry I.
- g_4 satisfies Reasons Symmetry II: Same reasoning as for Reasons Symmetry I.
- g_5 does not satisfy Reasons Symmetry II: Same reasoning as for Reasons Symmetry I.
- g_6 satisfies Reasons Symmetry II: Same reasoning as for Reasons Symmetry I.

Sixteenth, Menu Insensitivity

- g_1 does not satisfy Menu Insensitivity: Consider a Normative Graph N with one option o and a reason of positive weight against o . Then consider the Normative Graph N' containing the same reason and option, but also the additional option o' . Then $g(R_N) = \{o\}$ but $g(R_{N'}) = \{o'\}$.
- g_2 and g_3 satisfy Menu Insensitivity: Follows immediately from the definitions of g_2 and g_3 .
- g_4 does not satisfy Menu Insensitivity: Same proof as for g_1 , as long as we make sure, that the reason's weight is above the threshold l .
- g_5 satisfies Menu Insensitivity: Follows immediately from the definition of g_5 .
- g_6 does not satisfy Menu Insensitivity: Same proof as for g_1 .

□

Proposition 3.4.1. *Every rational weighting function that satisfies the following principles is a constant function assigning 0 to every consideration in any Normative Graph.*

1. *Modifier Scaling: Equality*
2. *No Valence Flips for Modifiers*

3. Strong Modifier Reciprocity

4. No Enablers

5. Attenuators as Undercutters

6. Atemporality

Proof. Let δ be an arbitrary consideration in some normative graph N . Let A_δ^+ be the set of its amplifiers, A_δ^- be the set of its attenuators and A_δ^u the set of its undercutters; we will show that $f_N^w(\delta) = 0$

As A_δ^+ is by definition finite and the weight of its elements are all in $\mathbb{Q}$, we can write it as a tuple of its weights without losing information. $(\frac{p_1}{q_1}, \frac{p_2}{q_2}, \dots, \frac{p_n}{q_n})$. Similarly, we write the tuple $(\frac{p'_1}{q'_1}, \dots, \frac{p'_m}{q'_m})$ for the weights of the attenuators. It is easy to see that $q_1 \dots q_n q'_1 \dots q'_m$ is a common denominator of all weights in either set.

By principle Modifier Scaling: Equality, we know that we can replace the first amplifier by $p_1 q_2 \dots q_n q'_1 \dots q'_m$ amplifiers of weight $\frac{1}{q_1 \dots q_n q'_1 \dots q'_m}$. Analogously, we replace all the other amplifiers and attenuators without changing the final weight of δ . We are left with a set of amplifiers and attenuators that all have the same denominator. Let us say we have k such amplifiers and l such attenuators. Either $k \geq l$ or $l \geq k$. In either case, thanks to principle Strong Modifier Reciprocity, we can eliminate the modifiers with the smaller number completely without changing the final weight.

We are left with a graph where either no modifiers are left or only amplifiers or only attenuators. Assume for now, without loss of generality, that it is the attenuators that disappear. Then we are left with $k - l$ amplifiers of weight $\frac{1}{q_1 \dots q_n q'_1 \dots q'_m}$, whose joint weight we will denote as y .

Applying principle Modifier Scaling: Equality again, we can replace these amplifiers by a single amplifier, δ' of weight y .

Next, using principle Attenuators as Undercutters, we know that there is a set of attenuators that we can add to the graph to make the final weight of the consideration 0. By the same kind of reasoning above, we can replace this set of attenuators by a single attenuator whose weight is the sum of their weights. Let this single attenuator have a weight of x .

We know (by the exact kind of reasoning as above) that we can treat the original graph as equivalent to the graph that has a single amplifier of weight $x + y$ and an attenuator of weight x (and potentially some undercutters), which is equivalent to having one amplifier

A. Appendix

of strength y and two attenuators of weight x and y respectively.

Now we know that, given the principle Atemporality, we can calculate the final weights of the graph in two steps, by first calculating the final weight of δ considering only the undercutters and the attenuator and then, using this final weight of δ as the new default weight and calculate the final weight considering only the amplifiers. Given principle No Valence Flips for Modifiers, the final weight of the graph containing only the attenuator of weight x is 0 (as it is already 0 even if we add back the original amplifiers). But treating this weight of 0 as the new default weight, by principle No Enablers, the final weight of the entire graph will also be 0. And hence, as the final weight of δ in the original graph is identical, the final weight of δ is 0.

Since the graph and the consideration were chosen arbitrarily, any rational function satisfying these six principles will always assign weight 0 to every consideration in every graph. $\square$

In order to prove Proposition 3.5.2, we first prove some lemmas.

Lemma A.3.1. *The consequence relation $\vdash_{f_1, g_1}$ is not identical to either*

- $\vdash_{f_2, g_1}$
- $\vdash_{f_3, g_1}$
- $\vdash_{f_4, g_1}$

Proof. Consider the following case (N_i, A_i)

- Let $N_1 = (\Delta_1, O_1)$ with $O_1 = \{o_1, o_2\}$ and $\Delta_1 = \{(p_1, o_1, r^+)\}$ with $A_1 = \{o_1\}$
- Let $N_2 = (\Delta_2, O_2)$ with $O_2 = \{o_1, o_2\}$ and $\Delta_2 = \{(p_1, o_1, r^+), (p_2, (p_1, o_1, r^+), a^-)\}$ with $A_2 = \{o_1, o_2\}$

With all three function combinations (f_2, g_1) , (f_3, g_1) , (f_4, g_1) these two cases are inconsistent (as an attenuator can never entirely take away the weight of a reason, unless it already starts out with a weight of 0) (which corresponds to the principle No Enablers).

With the function combination f_1, g_1 however, the two cases are not inconsistent, as can be seen for example by the weighting function w that assigns weight 3 to the reason (p_1, o_1, r^+) and weight 5 to the attenuator $(p_2, (p_1, o_1, r^+), a^-)$. Hence, the consequence relation for f_1, g_1 is different from the consequence relation for the other three function combinations. $\square$

Lemma A.3.2. *The consequence relation $\vdash_{f,g_1}$, for any $f \in \{f_1, \dots, f_4\}$, is not identical to either $\vdash_{f,g_2}$, $\vdash_{f,g_3}$, or $\vdash_{f,g_5}$.*

Proof. Consider the case (N_1, A_1) with $N_1 = (\Delta_1, O_1)$ and $O_1 = \{o_1\}$, with $r_1 = (p_1, o_1, r^-)$ and $\Delta_1 = \{r\}$ and $A = \emptyset$. This case is consistent for g_2, g_3 , and g_5 , as there exists a w that assigns r_1 a weight (larger than the threshold of for the inclusion of a reason in the case of g_5), such that $g(R_{N_1}) = \emptyset$ for any $g \in \{g_2, g_3, g_5\}$. But it is inconsistent for g_1 (as g_1 does not allow for an empty set of permissible actions) and so the consequence relations are different. Note that the choice of f is not relevant here, as for every f , the final weight will be equal to the default weight given by w .

□

Lemma A.3.3. *The consequence relation $\vdash_{f,g_1}$, for any $f \in \{f_1, \dots, f_4\}$ is not identical to $\vdash_{f,g_4}$ (for any choice $k > 0$ of the threshold for g_4).*

Proof. We first prove this for $f \neq f_1$. Consider the cases (N_2, A_2) and (N_3, A_3) where $N_2 = (\Delta_2, O_2)$ and $N_3 = (\Delta_3, O_2)$, such that $O_2 = \{o_2, o_3\}$ and $\Delta_2 = \{r_2 = (p_2, o_2, r^+), \delta_3 = (p_3, r_2, a^+)\}$ and $A_2 = \{o_2\}$. Let $\Delta_3 = \{r_2\}$ and $A_3 = O_2$. This is consistent for $\vdash_{f,g_4}$, as there is a w , such that $k > w(r_2) > 0$, but $f(r_2) > k$ and so $g_4(R_{N_2}) = \{o_2\}$ and $g(R_{N_3}) = O_2$. But in the case of g_1 , since $w(r_2) > 0$ and no other reason is part of the graph, $g_1(R_{N_3}) = \{o_2\} \neq O_2$. And so, if $f \neq f_1$, $\vdash_{f,g_4}$ is not identical to $\vdash_{f,g_1}$.

For the proof that $\vdash_{f_1,g_1}$ is not identical to $\vdash_{f_1,g_4}$, we will only give a rough sketch here, as the cases involved are quite numerous.

To do this, we need to make use of the same strategy we used in the proof of Proposition 3.5.3. Here, for g_4 and g_1 , we can create a series of cases that allows us to ensure that for three reasons r_1, r_2, r_3 for some option o , $w(r_2) = \frac{n-1}{n}w(r_1)$ for any $n \in \mathbb{N}$ and $w(r_3) = \frac{1}{n}w(r_1)$. This method only works if $w(r_3) > k > 0$, but as k is some fixed number, we can always create a set of cases that allows the three reasons to have this structure. Next, we introduce a further reason r^* such that $w(r_1) > w(r^*) > w(r_2)$. We can do this, by simply adding two cases where in the first the only reasons available are r_1 and r^* (for different options) and the option r_1 supports is the only permissible option. The second case is similar, only that here the option supported by r^* is the only permissible option.

Next, we can introduce an attenuator δ of r_1 with the weight $\frac{n-1}{n}w(r_1)$. We can achieve this by adding a case with r_1, δ and a reason with the same weight as r_3 for a different option.

A. Appendix

With some more cases, we can add a second attenuator δ' of r^* with the same weight as δ (we can, for example, do this, by first adding an amplifier of r^* , bringing it to the same (final) weight as r_1 , and then showing that given both the amplifier and the attenuator, the final weight is equal to r_3). We know that the final weight of $w(r^*)$ is not zero, but that it is smaller than $\frac{1}{n}w(r)$.

The last step is choosing the case such that in a two-choice case where there is only the reason r^* modified with the attenuator δ' , both options are allowed. For g_4 , this automatically implies that $w(r^*) < k$. But, for g_1 as it still has positive (final) weight, this combination of cases is impossible and so, finally, we reach the fact that given g_1 the combination of cases we created is inconsistent, but remains consistent for g_4 .

□

Lemma A.3.4. *The consequence relation $\vdash_{f,g_1}$, for any $f \in \{f_1, \dots, f_4\}$, is not identical to $\vdash_{f,g_6}$.*

Proof. We prove this statement one f at a time, starting with $f = f_1$. The main proof strategy we will use is making use of the fact that for a weighting function to satisfy a case, it has to satisfy a set of inequalities that are determined by the function governing the final weights f and the function governing the deontic statuses g . What we will do is provide a set of cases where for only one set of functions the (in)equalities have no solution, while giving an explicit solution for the other set.

First note, though, that we can always make sure that the (final or default) weight of a reason $r = (p, o, x)$ with $x \in \{r^+, r^-\}$ is positive by adding a case $((\Delta, O), A)$ with $\Delta = \{r\}$, $O = \{o, o'\}$, $\emptyset \subsetneq A \subsetneq O$.¹ Analogously, we can create a case that guarantees the final weight of r is positive, by adding all its modifiers (and their modifiers) and undercutters to Δ . Additionally, as we have seen in Section 3.5.2, we can always guarantee that any combination of reasons has the same weight as another reason; In particular, we can guarantee that two reasons have the same weight.

With this out of the way, we will start by proving the statement for $\vdash_{f_1,g_1}$. Let $w(r_1) = w(r_2) = w(r_3) > 0$, with r_1 being a reason for o and r_2, r_3 being reasons for o' . Now, consider the following case (N_1, A_1) with $N_1 = (\Delta_1, O)$ where $\Delta_1 = \{r_1, r_2, r_3, \delta = (p, r_1, a^+)\}$, $O = \{o, o'\}$ and $A_1 = O$. This means that, for g_6 (given f_1), $1 + w(\delta) + w(r_1) = (1 + w(r_1))^2$. And so, $w(\delta) = w(r_1)^2 + w(r_1)$. Let r_4 be yet another reason for o' with $w(r_4) = w(r_1)$ and $\delta' = (p', r_1, a^+)$ with $w(\delta') = w(\delta)$ (which we can guarantee

¹As long as the weighting function w allows for this. If not, we need to complicate the proof slightly and switch to a different function w^* that does allow for it, but keeps the weight of all other relevant considerations the same. We can then prove that the inferences licensed by w and w^* coincide for all the graphs we are interested in. We ignore this slight complication to keep the proof readable.

using the same method mentioned above). And the case (N_2, A_2) with $N_2 = (\Delta_2, O)$ with $\Delta_2 = \Delta_1 \cup \{r_4, \delta'\}$ and $A_2 = A_1 = O$. For g_6 (given f_1), this gives us the fact that $1 + (2 * w(\delta) + w(r_1)) = (1 + w(r_1))^3$. Combined with the earlier statement that $1 + w(\delta) + w(r_1) = (1 + w(r_1))^2$, this gives us, that $w(r_1) = 0$, contradicting our assumption. Thus, for g_6 (given f_1) the cases (N_1, A_1) and (N_2, A_2) together with the implicit cases we used to give the considerations the same weight (> 0) are inconsistent. Next, we show that these two cases are *not* inconsistent for g_1 (given f_1). The first case corresponds to the equality $w(\delta) + w(r_1) = 2 * w(r_1)$ and the second case corresponds to $2 * w(\delta) + w(r_1) = 3 * w(r_1)$. Clearly, there is some w that satisfies both constraints (and all the other ones we've implicitly assumed). For example, the function w' that assigns the constant weight 1 to every normative consideration mentioned.

Next we show that $\vdash_{f_2, g_1}$ is not identical to $\vdash_{f_2, g_6}$. We employ a very similar strategy. We will use the cases (N_3, A_3) and (N_4, A_4) , where $N_3 = (\Delta_3, O)$, $N_4 = (\Delta_4, O)$ and $A_3 = A_4 = O$.

We need one reason for o , r_1 with positive weight and two amplifiers of r_1 , δ_1 and δ_2 with equal weight. Additionally, we need four reasons for o' , r_2, r_3, r_4, r_5 where $w(r_1) = w(r_2) = w(r_3) = w(r_4) > 0$. We then define $\Delta_3 = \{r_1, \delta_1, r_3, r_4\}$ and $\Delta_4 = \{r_1, \delta_1, \delta_2, r_2, r_3, r_4, r_5\}$. Like before, these two cases give us (for g_6 given f_2) the following two equations: $(1 + (1 + w(\delta_1))w(r_1)) = (1 + w(r_1))^2$ and $(1 + (1 + w(\delta_1))^2 w(r_1)) = (1 + w(r_1))^4$. From the first equation, we can immediately derive that $w(\delta_1) = 1 + w(r_1)$. Plugging this in the second equation gives us that the only non-negative solution is $w(r) = 0$ contradicting our assumption. So given g_6 and f_2 , the two cases are inconsistent. They are however, not inconsistent given g_1 and f_2 as we will show next. For this combination of functions we obtain the following two functions. $(1 + w(\delta_1))w(r_1) = 2 * w(r_1)$ and $(1 + w(\delta_1))^2 w(r_1) = 4 * w(r_1)$. It is easy to see that if we choose $w(\delta_1) = 1$, every choice of $w(r)$ is a solution and hence the two cases are not inconsistent given g_1 and f_2 .

To show that $\vdash_{f_3, g_1}$ is not identical to $\vdash_{f_3, g_6}$ and $\vdash_{f_4, g_1}$ is not identical to $\vdash_{f_4, g_6}$ can reuse one case from before. Let (N_3, A_3) (where the set of considerations Δ was defined as $\Delta_3 = \{r_1, \delta_1, r_2, r_3\}$). We need additionally the case (N_5, A_5) with $N_5 = (\Delta_5, O)$ and $A_5 = O$ is constructed using the following additional amplifier $\delta'_1 = (p^*, r_1, a^+)$ and yet another reason of equal weight to r_1 for o' , r_6 such that $w(\delta'_1) = w(\delta_1)$. Then $\Delta_5 = \Delta_3 \cup \{\delta'_1, r_6\}$.

Next, note that, the (in)equalities generated in these cases for each g_1 and g_6 are the same no matter if one chooses f_3 or f_4 . We will show that the set of (in)equalities generated by g_6 (given either f_3 or f_4) has no positive solutions, and thus that both the cases are inconsistent for both $\vdash_{f_3, g_6}$ and $\vdash_{f_4, g_6}$. Then, in the next step, we will show that it has concrete solutions for $\vdash_{f_3, g_1}$ and $\vdash_{f_4, g_1}$. For the case of $\vdash_{f_3, g_6}$ (or $\vdash_{f_4, g_6}$), the (in)equalities are just the two equalities: $1 + ((1 + w(\delta_1))w(r_1)) = (1 + w(r_1))^2$ and $1 + ((1 + 2 * w(\delta_1))w(r_1)) = (1 + w(r_1))^3$. Again, this forces that $w(r_1) = 0$ and hence

A. Appendix

violates our assumption. This means that for both $f = f_3$ and $f = f_4$ together with the other implicit cases, the two cases are inconsistent given g_6 . We still need to show that they are consistent given g_1 , and here again, both choices of f produce the same equality that we need to show has a solution. They are: $(1 + w(\delta))w(r_1) = 2w(r_1)$ and $(1 + 2 * w(\delta))w(r_1) = 3w(r_1)$. Again, we can easily spot a solution to this, for example w' with $w'(r_1) = w'(\delta) = 1$. Hence, the cases (including the implicit cases) are consistent given g_1 and either f_3 or f_4 . This concludes our proof. $\square$

Proposition 3.5.2. *The consequence relation $\vdash_{f_1, g_1}$ is not identical to any consequence relation $\vdash_{f, g}$ with $f \in \{f_1, \dots, f_4\}$ and $g \in \{g_1, \dots, g_6\}$ where not both $f = f_1$ and $g = g_1$.*

Proof. The results follow immediately from Lemmas A.3.1, A.3.2, A.3.3, A.3.4. $\square$

A.4. Origin of the Material

Parts of this thesis result from published papers. Below, we list these papers and the contribution of each author.

Sections 2.1 - 2.4.2, Section 2.6.2 and part of Section 2.6.2 are based on the following paper.

Knoks, A., & Streit, D. (2023). Beyond reasons and obligations: A dual-role approach to reasons and supererogation. In *Deontic Logic and Normative Systems: 16th International Conference (DEON2023)* (pp. 119–137). London, United Kingdom: College Publications.

Both authors contributed equally to this paper.

Section 2.4.3 attempts to improve shortcomings of a paper presented at the American Philosophical Association Pacific Meeting.

Knoks, A., & Streit, D. (05 April 2023). *Supererogation, Dual-Role Views, and Defeasible Logic* [Paper presentation]. American Philosophical Association 2023 Pacific Division Meeting, San Francisco, United States - California.

Both authors contributed equally to this paper. The improvements in Section 2.4.3 were made by David Streit.

The idea of using a principle-based analysis to determine the final weights of reasons

in an argumentation framework (similar to Section 3.4) was presented (using a different model than the one found in this thesis) in the paper:

Streit, D., de Wit, V., & Knoks, A. (2023). Reasons in Weighted Argumentation Graphs. In N. Alechina (Ed.), *Logic, Rationality, and Interaction - 9th International Workshop, LORI 2023*, Proceedings (pp. 251–259). Springer Science and Business Media Deutschland GmbH.

The functions and principles to analyze were suggested by David Streit. For the remaining part of the paper, the authors contributed equally.

Chapter 4 is an extended version of the following paper:

Alcaraz, B., Knoks, A. & Streit, D. (2024). Estimating weights of reasons using metaheuristics: A hybrid approach to machine ethics, in: *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES2024)*, ACM Press.

The underlying model was created by David Streit. The main idea to use the model in a framework for machine learning was by Aleks Knoks, and (almost) all of the implementation in Python was done by Benoît Alcaraz. Otherwise, the authors contributed equally.

Bibliography

- D. Abel, J. Macglashan, and M. L. Littman. Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop on AI, Ethics, and Society*, 2016.
- C. Allen, I. Smit, and W. Wallach. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology*, 7:149–155, 2005.
- L. Amgoud and J. Ben-Naim. Evaluation of arguments from support relations: Axioms and semantics. In *25th International Joint Conference on Artificial Intelligence (IJCAI 2016)*, pages 900–906, 2016.
- L. Amgoud and J. Ben-Naim. Evaluation of arguments in weighted bipolar graphs. *International Journal of Approximate Reasoning*, 99:39–55, 2018.
- L. Amgoud, J. Ben-Naim, D. Doder, and S. Vesic. Acceptability semantics for weighted argumentation frameworks. In *Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)*. International Joint Conferences on Artificial Intelligence (IJCAI), 2017.
- L. Amgoud, D. Doder, and S. Vesic. Evaluation of argument strength in attack graphs: Foundations and semantics. *Artificial Intelligence*, 302:1–61, 2022.
- M. Anderson and S. L. Anderson. GenEth: A general ethical dilemma analyzer. *Paladyn, Journal of Behavioral Robotics*, 9(1):337–357, 2018.
- R. Arkin. *Governing Lethal Behavior in Autonomous Robots*. Chapman and Hall/CRC, 2009.
- G. Arrhenius, K. Bykvist, T. Campbell, and E. Finneron-Burns, editors. *The Oxford Handbook of Population Ethics*. Oxford University Press, 2022.
- K. J. Arrow. A difficulty in the concept of social welfare. *Journal of political economy*, 58(4):328–346, 1950.
- S. Artemov. The logic of justification. *The Review of Symbolic Logic*, 1(4):477–513, 2008.
- S. Artemov and M. Fitting. Justification Logic. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition, 2021.

Bibliography

- E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.
- E. Awad, M. Anderson, S. L. Anderson, and B. Liao. An approach for combining ethical principles with public opinion to guide public policy. *Artificial Intelligence*, 287:103349, 2020a. ISSN 0004-3702. doi: <https://doi.org/10.1016/j.artint.2020.103349>. URL <https://www.sciencedirect.com/science/article/pii/S0004370219301079>.
- E. Awad, S. Dsouza, A. Shariff, I. Rahwan, and J.-F. Bonnefon. Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, 117(5):2332–2337, 2020b. doi: 10.1073/pnas.1911517117. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1911517117>.
- R. Bader. Conditions, modifiers, and holism. In *Weighing Reasons*. Oxford University Press, 04 2016. ISBN 9780199315192. URL <https://doi.org/10.1093/acprof:oso/9780199315192.003.0002>.
- A. Balakrishnan, D. Bouneffouf, N. Mattei, and F. Rossi. Using multi-armed bandits to learn ethical priorities for online ai systems. *IBM Journal of Research and Development*, 63(4/5):1–1, 2019.
- P. Baroni and M. Giacomin. On principle-based evaluation of extension-based argumentation semantics. *Artificial Intelligence*, 171(10-15):675–700, 2007.
- M. Bedke. Passing the deontic buck. *Oxford Studies in Metaethcis*, 6:128–53, 2011.
- T. Bench-Capon, G. Robinson, T. Routen, and M. Sergot. Logic programming for large scale applications in law: a formalisation of supplementary Benefit Legislation. In *Proceedings of the First International Conference on Artificial Intelligence and Law (ICAIL-87)*, page 190–198, 1987.
- C. Benz Müller, X. Parent, and L. van der Torre. Designing normative theories for ethical and legal reasoning: LogiKEy framework, methodology, and tool support. *Artificial intelligence*, 287:103348, 2020.
- S. Berker. Particular reasons. *Ethics*, 118(1):109–139, 2007. doi: 10.1086/521586.
- S. Berker. A combinatorial argument against practical reasons for belief. *Analytic Philosophy*, 59(4):427–470, 2018. doi: 10.1111/phib.12140.
- P. Biro, D. F. Manlove, and R. Rizzi. Maximum weight cycle packing in directed graphs, with application to kidney exchange programs. *Discrete Mathematics, Algorithms and Applications*, 1(04):499–517, 2009.
- G. Boella and L. van der Torre. Institutions with a hierarchy of authorities in distributed dynamic environments. *Artificial Intelligence and Law*, 16:53–71, 2008.
- R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

- P. Bremner, L. A. Dennis, M. Fisher, and A. F. Winfield. On proactive, transparent, and verifiable ethical reasoning for robots. *Proceedings of the IEEE*, 107(3):541–561, 2019.
- G. Brewka. Reasoning about priorities in default logic. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, pages 940–5, 1994.
- I. Canavotto and J. Horty. Piecemeal knowledge acquisition for computational normative reasoning. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, AIES '22*, page 171–180, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471.
- J. Dancy. *Ethics without Principles*. Oxford University Press, 2004.
- J. Dancy. Moral Particularism. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2017 edition, 2017.
- F. Dietrich and C. List. A reason-based theory of rational choice. *Noûs*, 47(1):104–134, 2011. doi: 10.1111/j.1468-0068.2011.00840.x.
- F. Dietrich and C. List. What matters and how it matters: a choice-theoretic representation of moral theories. *Philosophical Review*, 126(4):421–479, 2017a.
- F. Dietrich and C. List. Choice-theoretic deontic logic. 2017b. Manuscript: <http://www.franzdietrich.net/Papers/DietrichList-ChoiceTheoreticDeonticLogic.pdf>.
- D. Dorsey. The supererogatory, and how to accommodate it. *Utilitas*, 25(3):355–382, 2013.
- D. Dorsey. *The Limits of Moral Authority*. Oxford University Press, 04 2016. ISBN 9780198728900. doi: 10.1093/acprof:oso/9780198728900.001.0001. URL <https://doi.org/10.1093/acprof:oso/9780198728900.001.0001>.
- D. Draï. Reasons have no weight. *Philosophical Quarterly*, 68(270):60–76, 2018. doi: 10.1093/pq/pqx025.
- F. Faroldi and T. Protopopescu. A hyperintensional logical framework for deontic reasons. *Logic Journal of the IGPL*, 27(4):411–433, 06 2019. ISSN 1367-0751. doi: 10.1093/jigpal/jzz012. URL <https://doi.org/10.1093/jigpal/jzz012>.
- F. Faroldi and T. Protopopescu. All-things-considered ought via reasons in justification logic. *Deontic Logic and Normative Systems*, page 257, 2023.
- M. Fischer and L. A. Dennis. *Verifiable Autonomous Systems*. Cambridge University Press, 2023.
- R. Freedman, J. S. Borg, W. Sinnott-Armstrong, J. P. Dickerson, and V. Conitzer. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence*, 283:103261, 2020.

Bibliography

- J.-G. Ganascia. Modelling ethical rules of lying with answer set programming. *Ethics and Information Technology*, 9:39–47, 2007.
- J. Gert. *Brute Rationality: Normativity and Human Action*. Oxford University Press, 2004.
- J. Gert. Normative strength and the balance of reasons. *Philosophical Review*, 116(4): 533–62, 2007.
- J. Gert. Moral worth, supererogation, and the justifying/requiring distinction. *Philosophical Review*, 121(4):611–8, 2011.
- J. Gert. Perform a justified option. *Utilitas*, 26(2):206–17, 2014.
- P. Greenspan. Asymmetrical practical reasons. In J. C. Marek and M. E. Reicher, editors, *Experience and Analysis: Proceedings of the 27th International Wittgenstein Symposium*, pages 387–94. Vienna: ÖBV and HPT, 2005.
- P. Greenspan. Practical reasons and moral ‘ought’. *Oxford Studies in Metaethcis*, 2: 181–205, 2007.
- J. Hansen. Reasoning about permission and obligation. *David Makinson on classical methods for non-classical problems*, pages 287–333, 2014.
- J. C. Harsanyi. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of political economy*, 63(4):309–321, 1955.
- J. Horton. The all or nothing problem. *Journal of Philosophy*, 114:94–104, 2017.
- J. F. Horty. Moral dilemmas and nonmonotonic logic. *Journal of Philosophical Logic*, 23:35–65, 1994.
- J. F. Horty. Reasoning with moral conflicts. *Noûs*, 37:557–605, 2003.
- J. F. Horty. *Reasons as Defaults*. New York: Oxford University Press, 2012.
- L. Jiang, J. D. Hwang, C. Bhagavatula, R. L. Bras, J. Liang, J. Dodge, K. Sakaguchi, M. Forbes, J. Borchardt, S. Gabriel, et al. Can machines learn morality? the Delphi Experiment. *arXiv preprint arXiv:2110.07574*, 2021.
- F. Kamm. Supererogation and obligation. *Journal of Philosophy*, 82:118–38, 1985.
- H. Kamp. Free choice permission. In *Meaning and the Dynamics of Interpretation*, pages 169–184. Brill, 2013.
- S. Kearns. Bearing the weight of reasons. *Weighing reasons*, pages 173–90, 2016.
- B. Kiesewetter. *The normativity of rationality*. Oxford University Press, 2017.

- A. Knoks and D. Streit. Beyond reasons and obligations: A dual-role approach to reasons and supererogation. In *Deontic Logic and Normative Systems: 16th International Conference (DEON2023)*. College Publications, London, United Kingdom, 2023.
- A. Kratzer. *Modals and Conditionals: New and Revised Perspectives*. Oxford: Oxford University Press, 2012.
- D. Lewis. *Counterfactuals*. Blackwell, 1973.
- H. Lin. Bayesian Epistemology. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2024 edition, 2024.
- E. Lord. Justifying partiality. *Ethical Theory and Moral Practice*, 19(3):569–590, 2016. doi: 10.1007/s10677-015-9675-4.
- E. Lord and B. Maguire. *Weighing Reasons*. Oxford University Press, 04 2016. ISBN 9780199315192. doi: 10.1093/acprof:oso/9780199315192.001.0001. URL <https://doi.org/10.1093/acprof:oso/9780199315192.001.0001>.
- D. Makinson and L. van der Torre. Permission from an input/output perspective. *Journal of philosophical logic*, 32:391–416, 2003.
- P. McNamara. Doing well enough: Toward a logic for common-sense morality. *Studia Logica*, 57:167–92, 1996a.
- P. McNamara. Must I do what I ought (or will the least I can do do)? In M. Brown and J. Carmo, editors, *Deontic Logic, Agency and Normative Systems*, pages 154–73. Springer-Verlag, 1996b.
- P. McNamara. Praise, blame, obligation, and DWE: Toward a framework for classical supererogation and kin. *Journal of Applied Logic*, 9:153–70, 2011.
- P. McNamara. Logics for supererogation and allied normative concepts. In D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, pages 155–306. College Publications, 2022a.
- P. McNamara. A natural conditionalization of the dwe framework. In *Agency, Norms, Inquiry, and Artifacts: Essays in Honor of Risto Hilpinen*, pages 113–136. Springer, 2022b.
- P. McNamara. Extensions and variations of the dwe framework with applications. *Journal of Applied Logics—IfCoLog Journal of Logics and their Applications*, 12:189–220, 2025.
- R. Mullins. Formalizing reasons, oughts, and requirements. *Ergo*, 7(22):568–99, 2021.
- D. Muñoz. From rights to prerogatives. *Philosophy and Phenomenological Research*, 102(3):608–623, 2020. doi: 10.1111/phpr.12682.

Bibliography

- D. Muñoz. Three paradoxes of supererogation. *Noûs*, 55(3):699–716, 2021. doi: 10.1111/nous.12326.
- D. Muñoz and T. Pummer. Supererogation and conditional obligation. *Philosophical Studies*, 179:1429–43, 2022.
- S. Nair. Conflicting reasons, unconflicting ‘oughts’. *Philosophical Studies*, 173:629–63, 2016.
- S. Nair. “adding up” reasons: Lessons for reductive and nonreductive approaches. *Ethics*, 132(1):38–88, 2021. doi: 10.1086/715288.
- S. Nofal, K. Atkinson, and P. E. Dunne. Computing grounded extensions of abstract argumentation frameworks. *The Computer Journal*, 64(1):54–63, 2021.
- M. Olszewski, X. Parent, and L. Van der Torre. Permissive and regulative norms in deontic logic. *Journal of Logic and Computation*, 34(4):728–763, 05 2023. ISSN 0955-792X. doi: 10.1093/logcom/exad024. URL <https://doi.org/10.1093/logcom/exad024>.
- D. Parfit. *On What Matters*. Oxford University Press, 2011.
- L. C. Paulson. Natural deduction as higher-order resolution. *The Journal of Logic Programming*, 3(3):237–258, 1986. ISSN 0743-1066. doi: [https://doi.org/10.1016/0743-1066\(86\)90015-4](https://doi.org/10.1016/0743-1066(86)90015-4). URL <https://www.sciencedirect.com/science/article/pii/0743106686900154>.
- J. Pearl. *Causality*. Cambridge university press, 2009.
- P. Portner. *Modality*. Oxford University Press, 2009.
- H. Prakken and G. Sartor. A dialectical model of assessing conflicting arguments in legal reasoning. *Logical models of legal argumentation*, pages 175–211, 1996.
- T. Pummer. *The Rules of Rescue: Cost, Distance, and Effective Altruism*. Oxford University Press, New York, 2023.
- W. Rabinowicz and T. Rønnow-Rasmussen. The strike of the demon: On fitting pro-attitudes and value. *Ethics*, 114(3):391–423, 2004.
- J. Raz. *Practical reason and norms*. Oxford University Press, 1990.
- J. Raz. *Practical Reason and Norms*. Oxford University Press, 09 1999. ISBN 9780198268345. doi: 10.1093/acprof:oso/9780198268345.001.0001. URL <https://doi.org/10.1093/acprof:oso/9780198268345.001.0001>.
- R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.
- S. Russell, D. Dewey, , and M. Tegmark. Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, 36(4):105–114, 2015.

- T. M. Scanlon. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press, 1998.
- T. M. Scanlon. *Being Realistic about Reasons*. Oxford University Press, 01 2014. ISBN 9780199678488. doi: 10.1093/acprof:oso/9780199678488.001.0001. URL <https://doi.org/10.1093/acprof:oso/9780199678488.001.0001>.
- S. Scheffler. Morality and reasonable partiality. In B. Feltham and J. Cottingham, editors, *Partiality and Impartiality: Morality, Special Relationships, and the Wider World*. Oxford University Press, 2010.
- M. Schroeder. *Reasons First*. Oxford University Press, 2021.
- M. Sergot, F. Sadri, R. Kowalski, F. Kriwaczek, P. Hammond, and H. T. Cory. The British Nationality Act as a logic program. In *Communications of the Association for Computing Machinery* 29, page 370–386, 1986.
- I. Sher. Comparative value and the weight of reasons. *Economics and Philosophy*, 35(1):103–158, 2019. doi: 10.1017/S0266267118000160.
- A. Silk. What normative terms mean and why it matters for ethical theory. *Oxford Studies in Normative Ethics*, 25:296–325, 2015.
- W. Sinnott-Armstrong, J. A. Skorburg, et al. How ai can aid bioethics. *Journal of Practical Ethics*, 9(1), 2021.
- J. Skorupski. *The domain of reasons*. Oxford University Press, 2010.
- H. M. Smith. Even more supererogatory. *Australasian Journal of Philosophy*, forthcoming. doi: 10.1080/00048402.2022.2153369.
- J. Snedegar. Reasons, oughts, and requirements. *Oxford Studies in Metaethcis*, 11: 155–81, 2016.
- J. Snedegar. *Contrastive Reasons*. Oxford University Press, New York, NY, 2017.
- J. Snedegar. Reasons for and reasons against. *Philosophical Studies*, 175(3):725–743, 2018. doi: 10.1007/s11098-017-0889-2.
- A. Stolpe. A theory of permission based on the notion of derogation. *Journal of applied logic*, 8(1):97–113, 2010.
- S. Stroud. Permissible partiality, projects, and plural agency. In B. Feltham and J. Cottingham, editors, *Partiality and Impartiality: Morality, Special Relationships, and the Wider World*. Oxford University Press, 2010.
- J. M. Taurek. Should the numbers count? *Philosophy & Public Affairs*, pages 293–316, 1977.

Bibliography

- C. Tucker. How to think about satisficing. *Philosophical Studies*, 174(6):1365–1384, 2017. doi: 10.1007/s11098-016-0763-7.
- C. Tucker. The dual scale model of weighing reasons. *Noûs*, 56(2):366–392, 2021. doi: 10.1111/nous.12361.
- C. Tucker. Parity, moral options, and the weights of reasons. *Noûs*, 57(2):454–480, 2022. doi: 10.1111/nous.12410.
- C. Tucker. A holist balance scale. *Journal of the American Philosophical Association*, 9(3):533–553, 2023. doi: 10.1017/apa.2022.13.
- C. Tucker. *The Weight of Reasons: A Framework for Ethics*. Oxford University Press, New York, forthcominga.
- C. Tucker. Weighing reasons against. *Oxford Studies in Metaethics*, forthcomingb.
- B. van Fraassen. Values and the heart’s command. *The Journal of Philosophy*, 70:5–19, 1973.
- K. von Fintel and S. Iatridou. Time and modality. In J. Guéron and J. Lecarme, editors, *How to say ought in foreign: The composition of weak necessity modals*, pages 115–41. Springer-Verlag, 2008.
- G. H. von Wright. *Norm and Action: A Logical Enquiry*. Routledge and Kegan Paul, New York, NY, USA, 1963.
- U. Wessels. Beyond the call of duty: The structure of a moral region. *Royal Institute of Philosophy Supplement: Supererogation*, 77:87–104, 2015.
- R. Wilson. Social choice theory without the pareto principle. *Journal of Economic Theory*, 5(3):478–486, 1972. ISSN 0022-0531. doi: [https://doi.org/10.1016/0022-0531\(72\)90051-8](https://doi.org/10.1016/0022-0531(72)90051-8).
- Y.-H. Wu and S.-D. Lin. A low-cost ethics shaping approach for designing reinforcement learning agents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.